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 e Decision Tree. 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 classificador usando 5,9 e 15 vizinhos para classificação obtiveram resultados melhores do que os demais. Além disso, esse parâmetros permitiram que um tempo razoável fosse gasto na classificação. Quanto a teste de hipótese, a algoritmo KNN para o valor de k-vizinhos igual a 7 obtém o primeiro lugar no ranking de modelos. Os demais algoritmos obtém um resultado similares entre si.

Base Bruta	mean	std	median	time
KNN3	0.958000	0.022716	0.96	0.004632
KNN5	0.959000	0.024880	0.96	0.004668
KNN7	0.964000	0.025508	0.97	0.004611
KNN9	0.964000	0.025768	0.97	0.004655
KNN15	0.963200	0.028103	0.96	0.004604
DecisionTree	0.959067	0.031227	0.96	0.004233

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 para diferentes valores de k. Nesse caso, os classificadores obtiveram um resultado similar de acordo com o teste de hipótese.

Base pré-processada	mean	std	median	time
KNN3	0.974000	0.015620	0.98	0.004351
KNN5	0.974000	0.014283	0.98	0.004338
KNN7	0.973333	0.016600	0.98	0.004386
KNN9	0.972000	0.018868	0.98	0.004362
KNN15	0.972800	0.019904	0.98	0.004361
DecisionTree	0.968000	0.025245	0.98	0.004018

2.2. Statlog

Em relação ao experimento usando a base de dados bruta, uma análise preliminar indica que o classificador com k igual a 3 tem um desempenho superior ao classificador usando outros valores para k-vizinhos, bem como supera de longe o algoritmo Decision Tree. Além de terem um uma acurácia média maior, ele apresenta resultados mais estáveis, tendo um menor desvio padrão. Além disso, eles gastam menos tempo. Portanto, com k igual 3 o classificador obtém um resultado superior os demais modelos. A análise do teste hipótese indica que o modelo Decision Tree apresentar um desempenho superior enquanto o Kneighbors com k igual 3 é o segundo do rank.

Em relação ao experimento usando a base de dados pré-processada, uma análise preliminar indica que o classificador com k=3 e 5 tem um desempenho superior ao classificador usando 15 e 9 vizinhos e ainda mais superior ao modelo Decision Tree. Além de terem um uma acurácia média maior, eles apresentam resultados mais estáveis, tendo

Base Bruta	Média	Desvio Padrão	Mediana	Tempo
KNN3	0.943644	0.006006	0.943644	0.041373
KNN5	0.937090	0.010135	0.936435	0.042068
KNN7	0.931717	0.012627	0.934469	0.042620
KNN9	0.927851	0.013484	0.926606	0.043175
KNN15	0.921232	0.018585	0.920708	0.044102
DecisionTree	0.926715	0.021239	0.926606	0.039589

Base Pré-processada	Média	Desvio Padrão	Mediana	Tempo
KNN3	0.831193	0.012417	0.833552	0.025606
KNN5	0.821232	0.015968	0.822412	0.025588
KNN7	0.812189	0.020659	0.813237	0.025904
KNN9	0.804522	0.023610	0.805374	0.026031
KNN15	0.794338	0.030978	0.799476	0.026349
DecisionTree	0.798550	0.030094	0.804718	0.023432

um menor desvio padrão. Além disso, eles gastam menos tempo. Portanto, com k=3 ou 5 o classificador obtém um resultado superior a quando k=15 e similar a quando k=5 e 7. Todos os algoritmos apresentaram um resultado inferior ao teste anterior.

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 igual a 9 e 15. O teste de hipótese confirma a análise inicial em que o KNN tem um resultado melhor com k igual a 15.

Base Bruta	Média	Desvio Padrão	Mediana	Tempo
KNN3	0.463091	0.007562	0.463636	0.055706
KNN5	0.466333	0.008440	0.468182	0.055990
KNN7	0.470586	0.010554	0.470909	0.056642
KNN9	0.474561	0.012490	0.473333	0.057304
KNN15	0.478352	0.014164	0.477576	0.058662
DecisionTree	0.471438	0.020781	0.473333	0.052180

Em um análise posterior, usando a base de dados pré-processada é possível ver que o algoritmo de classificação melhora de maneira substância para todos os valores de k e também para o algoritmo Decision Tree. Além disso, o melhor valor para k continua sendo k igual a 15 de acordo com o teste de hipótese e média obtida na tabela de resultados.

Base Pré-processada	Média	Desvio Padrão	Mediana	Tempo
KNN3	0.830848	0.006063	0.833030	0.052943
KNN5	0.833636	0.005690	0.833939	0.053423
KNN7	0.835657	0.005965	0.835758	0.054045
KNN9	0.837455	0.006280	0.837273	0.054640
KNN15	0.840194	0.008293	0.839394	0.055716
DecisionTree	0.835059	0.014327	0.837273	0.049198

3. Conclusão

Nesse trabalho, foi usado o algoritmo conhecido como k-Vizinhos mais próximos, (*KNeighbors Classifier*) ou KNN 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.