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 KNN. 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. 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 apenas 5 e 7 vizinhos para classificação obtive um resultado um pouco melhor do que os demais. Além disso, esse parâmetros permitiram que um tempo razoável fosse gasto na classificação. Quanto a teste de hipótese, como a hipótese nula não foi rejeita, podemos dizer que os classificadores obtiveram resultados equivalentes em termos de desempenho.

Base Bruta	k	Média	Desvio Padrão	Mediana	Tempo
0	3	0.963333	0.034801	0.966667	0.004348
1	5	0.966667	0.031623	0.966667	0.004199
2	7	0.967778	0.029165	0.966667	0.004193
3	9	0.966667	0.029814	0.966667	0.004177
4	15	0.966667	0.029059	0.966667	0.004170

Table 1. A tabela mostra o resultado da acurácia para os diferentes valores de k.

No segundo experimento, foi realizado o experimento usando as base de dados pré-processadas. Nesse experimento, fica claro um melhora significativa na média de desempenho do classificador KNN para diferentes valores de k. Nesse caso também a hipótese nula deixou de ser rejeitada e os classificadores obtiveram um resultado similar.

Base Pré-processada	k	Média	Desvio Padrão	Mediana	Tempo
0	3	0.980000	0.022111	0.983333	0.004011
1	5	0.980000	0.022111	0.983333	0.003885
2	7	0.981111	0.022250	1.000000	0.003993
3	9	0.980000	0.025604	1.000000	0.004005
4	15	0.980667	0.025024	1.000000	0.004048

2.2. Statlog

Base Bruta	k	Média	Desvio Padrão	Mediana	Tempo
0	3	0.948052	0.012509	0.946970	0.031639
1	5	0.942857	0.014068	0.943723	0.031883
2	7	0.938961	0.015391	0.939394	0.032150
3	9	0.935444	0.016681	0.933983	0.032513
4	15	0.929870	0.019629	0.929654	0.033240

Table 2. A tabela mostra o resultado da acurácia para os diferentes valores de k.

Em relação ao experimento usando a base de dados bruta, uma análise preliminar indica que o classificador com k=3 e 5 tem um desempenho um pouco superior ao classificador usando 15 e 9 vizinhos. O desempenho se mostra também um pouco superior aos demais tanto em relação a acurácia quanto em relação ao tempo gasto. Portanto, com k=3 ou 5 o classificador obtém um resultado superior a quando k=15 e similar a quando k=5 e 7. A análise do teste hipótese a seguir indicar que o KNeighbohrs apresentar um desempenho superior com k=3 sendo o 1 do rank.

No segundo experimento, o resultado do classificador para todos os valores foi inferior ao do primeiro experimento. Ainda assim o classificador continua obtendo os melhores resutlados para k=3 e a 5. O teste de hipótese confirma mais uma vez o resultado apontando no teste anterior.

Base Pré-processada	k	Média	Desvio Padrão	Mediana	Tempo
0	3	0.844156	0.017691	0.849567	0.017493
1	5	0.837013	0.017929	0.840909	0.017423
2	7	0.828644	0.021811	0.836580	0.017455
3	9	0.820942	0.025399	0.819264	0.017558
4	15	0.810390	0.031885	0.810606	0.017779

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. K-3 tem um menor tempo porém também tem o pior resultado. O teste de hipótese confirma a análise inicial em que o KNN tem um resultado melhor com k=15.

Base Bruta	k	Média	Desvio Padrão	Mediana	Tempo
0	3	0.462900	0.012661	0.4640	0.036819
1	5	0.464750	0.010954	0.4645	0.037420
2	7	0.471200	0.015003	0.4700	0.037948
3	9	0.475275	0.016162	0.4740	0.038510
4	15	0.479080	0.018015	0.4765	0.039376

Table 3. A tabela mostra o resultado da acurácia para os diferentes valores de k.

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. Além disso, o melhor valor para k continua sendo k=15 de acordo com o teste de hipótese e média obtida na tabela de resultados.

Base Pré-processada	k	Média	Desvio Padrão	Mediana	Tempo
0	3	0.822900	0.008769	0.8190	0.036294
1	5	0.829650	0.011693	0.8310	0.036157
2	7	0.832867	0.012452	0.8335	0.037180
3	9	0.835725	0.013155	0.8370	0.037668
4	15	0.838700	0.013874	0.8380	0.038148

4. 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.