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, Decision Tree e Multinomial Naïve Bayes (NB). 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, o algoritmo KNN para todos os valores K obtiveram resultados similares entre si, estando nas

primeiras posições. Os demais, Decision Tree e MultinomialNB estiveram nas úlitmas posições.

Base Bruta classifier_name	Média	Mediana	Desvio Padrão	Tempo Médio
DecisionTree	0.9368	0.94	0.035077	0.002700
KNN-15	0.9600	0.96	0.036140	0.005564
KNN-3	0.9580	0.96	0.022946	0.005129
KNN-5	0.9600	0.96	0.027105	0.005312
KNN-7	0.9740	0.98	0.023990	0.005504
KNN-9	0.9640	0.97	0.026802	0.005710
MultinomialNB	0.7680	0.79	0.182991	0.003159

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, 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.

Base Pré-processada classifier_name	Média	Mediana	Desvio Padrão	Tempo Médio
DecisionTree	0.941333	0.94	0.032906	0.002605
KNN-15	0.970667	0.98	0.029187	0.005630
KNN-3	0.968667	0.98	0.019854	0.005205
KNN-5	0.969333	0.98	0.019888	0.005252
KNN-7	0.972667	0.98	0.021665	0.005265
KNN-9	0.966667	0.97	0.025028	0.005374
MultinomialNB	0.618667	0.61	0.184058	0.003102

2.2. Statlog

Base Bruta classifier_name	Média	Mediana	Desvio Padrão	Tempo Médio
DecisionTree	0.953578	0.951507	0.008403	0.017438
KNN-15	0.894758	0.897117	0.011235	0.051729
KNN-3	0.943644	0.943644	0.006067	0.046977
KNN-5	0.930537	0.931193	0.009230	0.047558
KNN-7	0.920970	0.920708	0.010084	0.048528
KNN-9	0.916252	0.916776	0.008424	0.050789
MultinomialNB	0.000000	0.000000	0.000000	0.002662

Em relação ao experimento usando a base de dados bruta, uma análise preliminar indica que o classificador Decision Tree obteve um melhor resultado em relação aos demais. Além de ter uma acurácia média maior, ele apresenta resultados mais estáveis, tendo um dos menores desvios padrão. Além disso, ele gastam 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	0.817588	0.815203	0.008860	0.009002
KNN-15	0.784404	0.789646	0.020699	0.031447
KNN-3	0.843644	0.847969	0.014743	0.029218
KNN-5	0.830668	0.834207	0.016470	0.029751
KNN-7	0.819528	0.823067	0.016430	0.030326
KNN-9	0.810747	0.807339	0.018480	0.030373
MultinomialNB	0.415334	0.422674	0.040080	0.003791

Em relação ao experimento usando a base de dados pré-processada, uma análise preliminar indica que o classificador com k igual 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 um menor desvio padrão. Além disso, eles gastam menos tempo. Portanto, com k igual 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. 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 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.

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 o algoritmo Decision Tree. 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 MultinomialNB e Decision Tree tem os melhores resultados.

3. Conclusão

Nesse trabalho, foi usado o algoritmo conhecido como k-Vizinhos mais próximos, (*KNeighbors Classifier*) ou KNN Decision Tree e MultinomialNB para classificar um

Base bruta classifier_name	Média	Mediana	Desvio Padrão	Tempo Médio
DecisionTree	0.437018	0.441515	0.012274	0.020108
KNN-15	0.493515	0.490303	0.009672	0.071407
KNN-3	0.463091	0.463636	0.007639	0.062101
KNN-5	0.469576	0.470909	0.008098	0.063875
KNN-7	0.479091	0.479091	0.009213	0.065249
KNN-9	0.486485	0.483030	0.010116	0.066456
MultinomialNB	0.000000	0.000000	0.000000	0.002342

Pré-proessado classifier_name	Média	Mediana	Desvio Padrão	Tempo Médio
DecisionTree	0.330511	0.333552	0.013887	0.015366
KNN-15	0.313893	0.311271	0.013787	0.032512
KNN-3	0.298296	0.300131	0.006651	0.029406
KNN-5	0.301573	0.301442	0.011178	0.029805
KNN-7	0.300524	0.303408	0.010492	0.030383
KNN-9	0.314548	0.316514	0.011343	0.031031
MultinomialNB	0.353997	0.353866	0.010211	0.003690

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.