Segundo trabalho de inteligencia artificial

Eduardo Montagner de Moraes Sarmento

Abstract

Este trabalho foi feito para comparar o desempenho em termos de acurácia de vários classificadores diferentes para um conjunto predeterminado de bases de dados, sendo os classificadores escolhidos o ZeroR, OneR, OneR Probabilístico, Centroide, Centroide OneR, Naive Bayes Gaussiano, Knn, Árvore de Decisão, Rede Neural e Florestas de Árvores.

Eles foram separados entre métodos que necessitam de ajuste de hiperparâmetros dos que não precisão e então foram testados, no caso dos que não precisam foi feita a validação cruzada com 10 folds e os que precisam foi feita ciclos de treino validação teste com 4 folds no ciclo interno e 10 folds no ciclo externo, e então obtivemos a média e o desvio padrão desses resultados e essas métricas foram utilizadas na comparação dos classificadores.

Ao final deste trabalho percebemos o quão difícil a tarefa de classificação é, sendo realmente necessário o uso de técnicas de classificação mais rebuscadas, como florestas de arvores que foi o classificador que se saiu melhor nos testes.

1. Introdução

Este trabalho consistiu em realizar uma comparação experimental entre um conjunto pré-definido de técnicas de aprendizado e classificação automática aplicadas a alguns problemas de classificação. As técnicas escolhidas foram: ZeroR, OneR, OneR Probabilístico, Centroide, Centroide OneR, Naive Bayes Gaussiano, Knn, Árvore de Decisão, Rede Neural e Florestas de Árvores. As bases de dados utilizadas foram iris, digits, wine e breast cancer, vindas do pacote datasets da biblioteca scikit-learn.

Para cada base, o procedimento experimental foi dividido em duas etapas. A primeira etapa consistiu no treino e teste com validação cruzada de 10 folds dos classificadores que não possuem hiper-parâmetros, isto é, os classificadores ZeroR, OneR, OneR Probabilístico, Centroide, Centroide OneR e Naive Bayes Gaussiano. Os resultados de cada classificador seram apresentados posteriormente numa tabela contendo a média das acurácias e o desvio padrão dos resultados obtidos em cada fold, e também através do boxplot dos resultados de cada classificador em cada fold.

A segunda etapa consistiu no treino, validação e teste dos classificadores que precisam de ajuste de hiperparâmetros, isto é, os classificadores Knn, Árvore de Decisão, Redes Neurais e Florestas de Árvores. Neste caso o procedimento de treinamento, validação e teste foi realizado através de ciclos aninhados de validação e teste, com o ciclo interno de validação contendo 4 folds e o externo de teste com 10 folds.

2. Descrição dos Datasets

As bases de dados utilizadas foram iris, digits, wine e breast cancer, vindas do pacote datasets da biblioteca scikit-learn.

Iris é talvez a mais famosa base de dados na literatura de reconhecimento de padrões, ela contem 3 classes e 50 instancias de cada classe, sendo que as classes se referem a 3 especies da planta iris, nela contem características como a tamanho da pétala e da sépala da flor da planta, tanto em largura como em comprimento, e com isso se pode classificar a qual especie aquela planta se encaixa.

Digits é uma base de dados que contem 1797 imagens 8x8 de digits escritos a mão por 44 escritores diferentes, as imagens são representadas como vetores de 64 posições em que cada posição determina se um pixel esta ligado ou não na tela, e com isso podemos classificar qual digito aquele conjunto de pixels forma.

Wine é uma base de dados que contem o resultado da analise química de vinhos vindos da mesma região da Itália, mas de 3 viniculturas diferentes, essa

analise nos da 13 caracteristicas quimicas de cada vinho que são utilizadas para classificar os vinhos.

Por ultimo temos a base de dados breast cancer que contem 699 casos clínicos de câncer de mama, cada caso tem 11 características que são utilizadas para a classificação.

3. Descrição dos Métodos Implementados

Nesta seção vamos descrever brevemente os métodos que foram implementados para esse trabalho, no caso os métodos ZeroR, OneR, OneR Probabilístico, Centroide e Centroide OneR, descrevendo a ideia por trás deles, como eles são treinados e como eles classificão novos dados

3.1. ZeroR

O mais simples dos classificadores, o ZeroR classifica sempre na classe majoritária, por isso ZeroR pois ele não tem nenhuma regra de classificação, com isso o treinamento dele é simplesmente determinar qual é a classe majoritária, que sera então usada para classificar todos os dados futuros.

3.2. OneR

Um pouco mais complicado que o ZeroR, o classificador OneR usa a característica que tem mais correlação com a classe dos dados para criar regras que serão utilizadas na classificação, mas ele só trabalha com dados discretizados. Sendo assim o processo de treinamento do classificador consistem em discretizar os dados de treino e então usando tabelas de contingencia determinar qual característica tem a maior correlação com a classe dos dados, e então essa característica é utilizada para fazer as regras de classificação, para isso é utilizado a tabela de contingencia dela novamente, em que para cada valor da característica é determinado qual classe ela tem o maior numero de dados com aquele valor para aquela classe e então qualquer dado futuro que sera classificado que tenha aquele valor sera classificado com aquela classe.

3.3. OneR Probabilístico

Este classificador se é quase igual ao OneR não probabilístico, ele também tenta determinar qual característica sozinha mais influencia na classificação do dado, e até o momento da classificação em si ele faz exatamente os mesmos passos, mas na hora de classificar ele leva em conta que existe a probabilidade de que os que seriam escolhidos no oneR normal para serem os valores das regras podem estar errados, então na hora de classificar ele aleatoriza o valor, segundo o peso dele, para tentar classificar corretamente tambem os que não seguiriam a regra do valor majoritario.

3.4. Centroide

O classificador centroide utiliza as distancias entre os pontos que os dados representam no espaço de solução para determinar a qual classe aquele dado pertence, para isso os dados são agrupados e classificados segundo centroides, que são pontos equidistantes de todos os outros dados daquela classe, então o processo de treinamento do classificador centroide consiste em achar os centroides dos dados passados para ele para que ele possa utilizar a distancia desses centroides a novos dados e então classifica-los.

3.5. Centroide OneR

O classificador Centroide OneR mistura os classificadores centroide e oner, nele primeiro ha os passos do oner para determinar qual característica mais influencia na classificação, então é feito os passos do centroide considerando apenas essa característica.

4. Descrição dos Experimentos Realizados

Nesta seção descreveremos os experimentos realizados e apresentaremos os resultados alcançados, também analisaremos estes dados.

Para os classificadores que necessitam de ajuste de hiper-parâmetros foram considerados os seguintes valores de hiper-parâmetros no gridsearch:

- Knn: [n-neighbors = 1, 3, 5, 7, 10]
- Arvore de Decisão: [max-depth = None, 3, 5, 10]
- Rede Neural: [max-iter = 50, 100, 200], [hidden-layer-sizes=(15)]
- Florestas de Arvores: [n-estimators = 10, 20, 50, 100]

4.1. Iris

Começamos os experimentos com a base de dados Iris e realizamos os experimentos primeiramente com os classificadores que não necessitam de ajuste de hiper-parâmetros, a tabela 1 mostra as médias e desvios padrões das acurácias dos classificadores em 10 folds de cross validation. Nela podemos ver que todos

Algoritmos	Média	Desvio padrão
ZeroR	0.3333333333333333	0.0
OneR	0.9533333333333334	0.07062332703142533
OneR Probabilístico	0.9	0.07200822998230953
Centroide	0.9333333333333333	0.06285393610547088
OneR Centroide	0.96	0.05621826951410451
Naive Bayes	0.9533333333333334	0.04499657051403685

Table 1: Tabela de resultados da primeira parte dos experimentos para o dataset Iris

os classificadores ficaram quase iguais em média para a classificação na base iris, menos o zeroR que teve uma média de 0.33 de accuracia, muito inferior aos outros, mas no seu desvio padrão foi 0, ou seja ele teve a mesma acurácia para todos os folds, e o oneR probabilístico teve uma média de 0.9, menor que os outros mas nem tanto quanto o zeroR, os outros tiveram a média de 0.953, sendo a média do centroide um pouco menor, mas o que obteve a melhor média foi o oneR centroide, mesmo que com uma margem pequena com relação ao naive bayes e ao oneR.

No boxplot 1 podemos ver esse resultado, o zeroR teve todas as suas previsões com a mesma acurácia por isso sua caixa é tão fina, enquanto os outros tiveram

resultados variados, o oneR obteve resultados bons mas teve um outlier ruim o que reduziu sua média e aumentou sua variancia, o oneR probabilístico obteve resultados piores do que o oneR comum e do que o oneR centroide, o centroide teve resultados bons e pouca variancia mas teve um outlier ruim e seus resultados não chegaram ao mesmo maximo que os outros por isso sua média levemente pior, ja o naive bayes e o oneR centroide obtiveram praticamente os mesmos resultados mas a média do oneR centroide foi ligeiramente melhor, no quesito de variância os dois foram praticamente iguais sendo o oneR centroide ligeiramente maior na variância, portanto consideramos que ambos os classificadores são igualmente bons para a base de dados iris e são os melhores para essa base de dados.

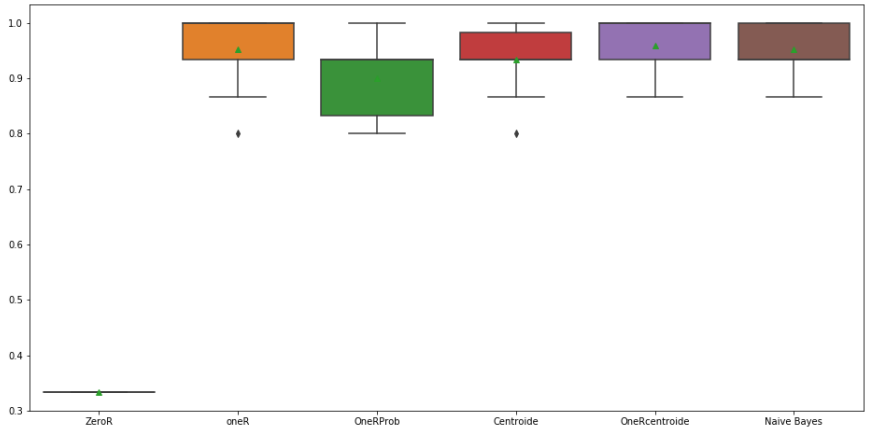

Figure 1: Boxplot da primeira parte dos experimentos para o dataset Iris

Agora para a segunda parte dos experimentos para a base de dados iris, a tabela 2 mostra as médias e desvios padrões das acurácias, e os hiper-parâmetros selecionados dos classificadores em 10 folds do ciclo externo de treino e teste e 4 folds do ciclo interno de treino e validação.

Nela podemos ver que os classificadores baseados em arvores foram melhores, no caso o de arvore de decisão e floresta de arvores, mas o classificador knn não

Algoritmos	Média	Desvio padrão	Hiper-parâmetro
KNN	0.9533333333333334	0.06324555320336757	10
Arvore de decisão	0.9666666666666667	0.03513641844631532	None
Rede neural	0.78	0.19385625322279457	15,200
Floresta de arvores	0.96	0.04661372658534006	100

Table 2: Tabela de resultados da segunda parte dos experimentos para o dataset Iris

ficou muito atrás, ele apenas foi um pouco pior tendo a média de acurácias de 0.95 enquanto os baseados em arvore ficaram próximos de 0,96, sendo assim eles esses classificadores ficaram com médias próximas, já o rede neural teve um resultado pior do que os outros, chegando a ficar com média de 0.78. Em desvio padrão o que teve o menor foi a arvore de decisão, enquanto o classificador rede neural teve o maior desvio padrão.

A figura 2 mostra o boxplot dos classificadores testados para a base de dados iris.

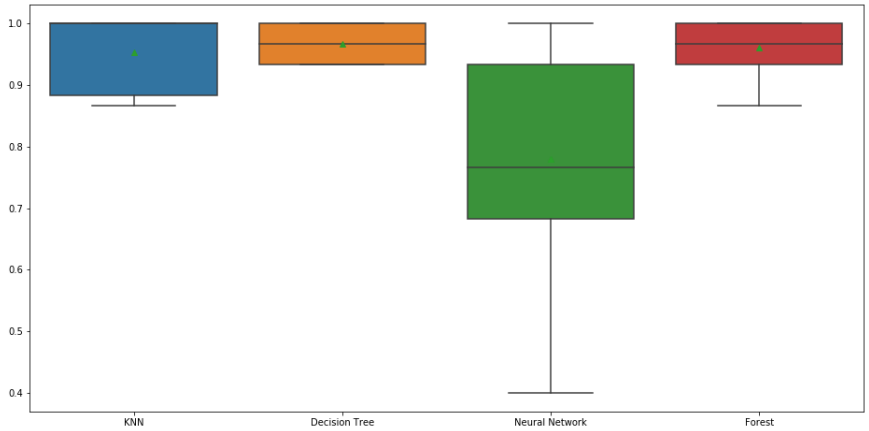

Figure 2: Boxplot da segunda parte dos experimentos para o dataset Iris

Nela nós podemos ver que realmente o rede neural foi o pior em média e em variância, enquanto os outros tiveram médias próximas, mas o arvore de decisão

foi o que teve a menor variância com uma média levemente superior aos outros e portanto ele foi considerado como o melhor para a base de dados iris.

4.2. Digits

Em seguida realizamos os experimentos na base digits com os classificadores que não necessitam de ajuste de hiper-parâmetros, a tabela 3 mostra as médias e desvios padrões das acurácias dos classificadores em 10 folds de cross validation.

Algoritmos	Média	Desvio padrão
ZeroR	0.10127425688130931	0.0013433375182975277
OneR	0.09905411468374282	0.0018502681832288986
OneR Probabilístico	0.09905411468374282	0.0018502681832288986
Centroide	0.8836101717889818	0.04335144121070077
OneR Centroide	0.12136205266868794	0.02351793058538051
Naive Bayes	0.8103537583567821	0.05972003743201967

Table 3: Tabela de resultados da primeira parte dos experimentos para o dataset Digits

Nela podemos ver que para essa base de dados os classificadores simples, zeroR, oneR e oneR probabilistico, tiveram resultados péssimos, em torno de 0.10 de acurácia, enquanto os classificadores mais complexos tiveram resultados bons com acurácias na faixa de 0.8, menos o oneRcentroide que mesmo sendo mais complexo teve resultado péssimo de 0.12 de média de acurácia, sendo o centroide na média o melhor entre eles. No caso do desvio padrão os mais simples tiveram desvios padrões menores e portanto tiveram menos variância nas acurácias.

A figura 3 mostra o boxplot dos classificadores testados na primeira parte dos experimentos para a base de dados digits.

Nela podemos visualizar as informações da tabela 3, os classificadores mais simples tiveram resultados péssimos e com pouca variância, mostrado pela finura e altura de suas caixas, enquanto os mais complexos tiveram variância maior mas médias muito melhores, sendo assim consideramos que o classificador centroide

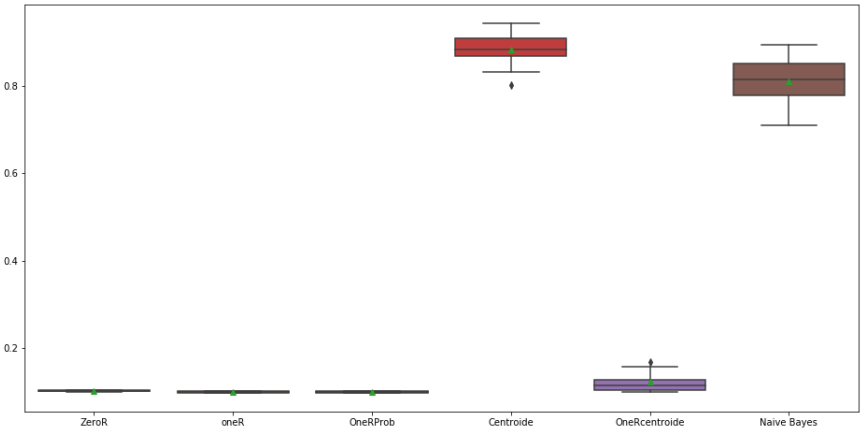

Figure 3: Boxplot da primeira parte dos experimentos para o dataset Digits

foi o melhor para essa base de dados na primeira fase dos experimentos pela sua média maior e baixo desvio padrão.

Agora para a segunda parte dos experimentos para a base de dados digits, a tabela 4 mostra as médias e desvios padrões das acurácias, e os hiper-parâmetros selecionados dos classificadores em 10 folds do ciclo externo de treino e teste e 4 folds do ciclo interno de treino e validação.

Algoritmos	Média	Desvio padrão	Hiper-parâmetro
KNN	0.96	0.04661372658534006	1
Arvore de decisão	0.8236750883726512	0.03627784675755085	10
Rede neural	0.9298402244124375	0.03165245227424336	15,200
Floresta de arvores	0.9472563484949691	0.027189718657094682	50

Table 4: Tabela de resultados da segunda parte dos experimentos para o dataset Digits

Nela podemos ver que os classificadores ficaram com médias boas para essa base de dados, nenhum abaixo dos 0.8 de acurácia em média ao contrario da primeira parte do experimento, mas tivemos uma diferença de médias grande também, sendo o classificador arvore de decisão o com a menor média com

acurácia de 0.82 enquanto o segundo menor teve acurácia de 0.92, sendo este a rede neural que teve desempenho melhor nessa base de dados do que na iris, os dois melhores tiveram médias próximas, o knn tendo a melhor média mas o floresta de arvores teve o menor desvio padrão tendo média similar.

A figura 4 mostra o boxplot dos classificadores testados na segunda parte dos experimentos para a base de dados digits.

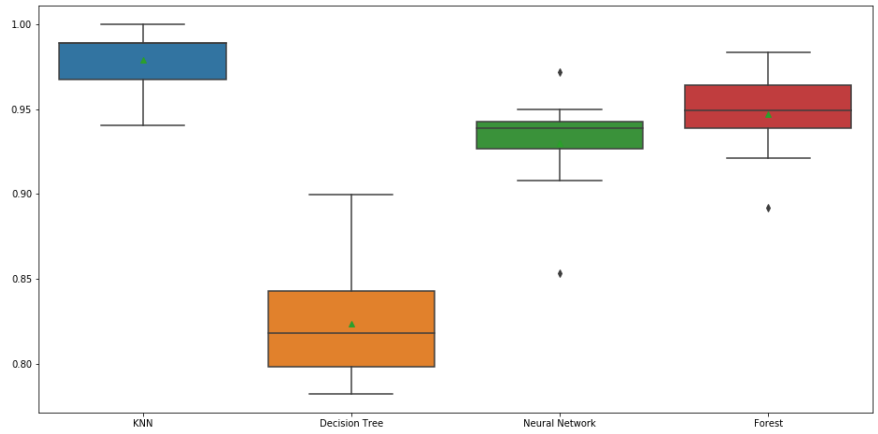

Figure 4: Boxplot da segunda parte dos experimentos para o dataset Digits

Nele podemos visualizar o que estava na tabela 4, o classificador arvore de decisão foi o pior de todos enquanto os outros se mantiveram próximos em com a média de mais de 0.9, mas aqui podemos ver que tanto o rede neural quanto o floresta tiveram outliers ruins, enquanto o knn não tem outliers concentrando todos os seus resultados entre 0.95 e 1, sendo assim ele foi considerado como o melhor classificador para esta base de dados.

4.3. Wine

Então seguida realizamos os experimentos na base wine com os classificadores que não necessitam de ajuste de hiper-parâmetros, a tabela 5 mostra as médias e

desvios padrões das acurácias dos classificadores em 10 folds de cross validation.

Algoritmos	Média	Desvio padrão
ZeroR	0.3992539559683522	0.01778421058376654
OneR	0.6640436876504988	0.06398265651745332
OneR Probabilístico	0.618468352253182	0.09059491442448983
Centroide	0.7216073271413829	0.08949384903682145
OneR Centroide	0.7216073271413829	0.08949384903682145
Naive Bayes	0.9616959064327485	0.04473779764296964

Table 5: Tabela de resultados da primeira parte dos experimentos para o dataset Wine

Essa base teve uma grande diferença entre as médias de todos os classificadores,indo do zeroR com 0.39 de média até o naive bayes com 0.96 de média, os classificadores oneR obtiveram médias similares mas não boas, estando na casa do 0.60, menos o oneR centroide que teve a mesma média do centroide de 0.72, e o naive bayes teve a melhor média, com a média mais de 0.20 acima dos segundos melhores, o centroide e oneR centroide, e mesmo no desvio padrão ele teve o segundo menor desvio padrão, perdendo apenas para o zeroR.

A figura 5 mostra o boxplot dos classificadores testados na primeira parte dos experimentos para a base de dados wine.

Nela podemos ver que o naive bayes realmente foi o melhor dentre todos os classificadores, tendo a melhor média, pouco desvio padrão e não tendo nenhuma acurácia inferior ao de outro classificador.

Agora para a segunda parte dos experimentos para a base de dados wine, a tabela 6 mostra as médias e desvios padrões das acurácias, e os hiper-parâmetros selecionados dos classificadores em 10 folds do ciclo externo de treino e teste e 4 folds do ciclo interno de treino e validação.

Nela podemos ver que assim como na primeira parte houve muita variação entre as médias dos classificadores, sendo que o primeiro colocado em média e em desvio padrão o classificador floresta de arvores, com uma acurácia de 0.97

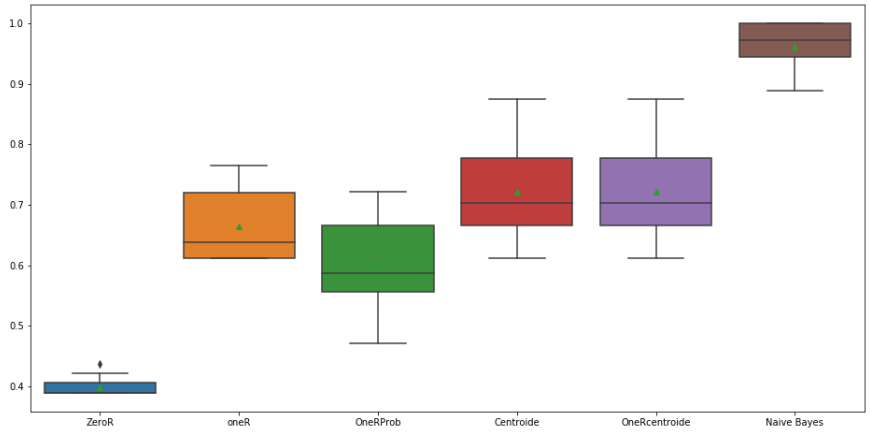

Figure 5: Boxplot da primeira parte dos experimentos para o dataset Wine

Algoritmos	Média	Desvio padrão	Hiper-parâmetro
KNN	0.7089245786033712	0.13101046785267526	1
Arvore de decisão	0.8891468868249054	0.05825738093935165	None
Rede neural	0.5515415376676986	0.16206756627820595	15,50
Floresta de arvores	0.9780701754385965	0.028323861767486157	10

Table 6: Tabela de resultados da segunda parte dos experimentos para o dataset Wine

na média quase 0.1 a mais do que o segundo colocado, o classificador arvore de decisão, além disso vemos que novamente o classificador rede neural foi o pior dentre os classificadores testados, tanto em média quanto em desvio padrão.

A figura 6 mostra o boxplot dos classificadores testados para a base de dados wine.

Nele podemos ver que o floresta de arvores realmente foi o melhor dentre os classificadores, tendo a melhor média dentre todos com a menor variância, além disso todas as suas acurácias foram maiores do que qualquer uma dos outros classificadores.

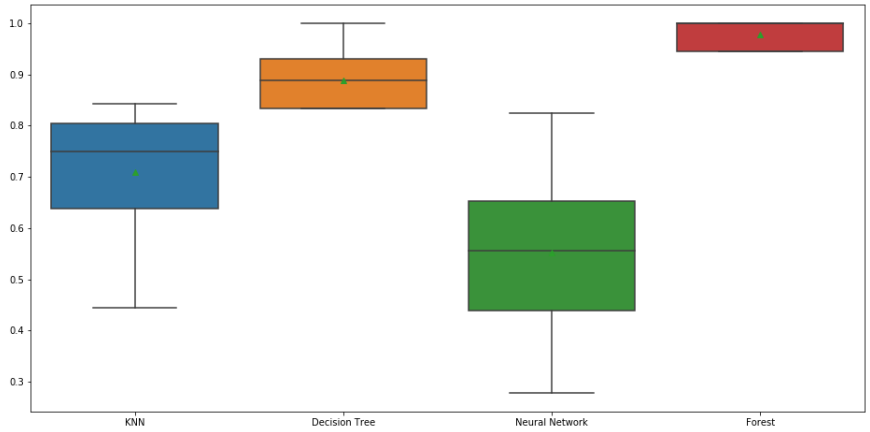

Figure 6: Boxplot da segunda parte dos experimentos para o dataset Wine

4.4. Breast Cancer

Por ultimo realizamos os experimentos na base breast cancer com os classificadores que não necessitam de ajuste de hiper-parâmetros, a tabela 7 mostra as médias e desvios padrões das acurácias dos classificadores em 10 folds de cross validation.

Algoritmos	Média	Desvio padrão
ZeroR	0.6274274047186933	0.004650542988181578
OneR	0.3725725952813067	0.004650542988181606
OneR Probabilístico	0.4022178290553971	0.04088347137294354
Centroide	0.8913641863278887	0.04089229703462307
OneR Centroide	0.66236388384755	0.0736920498357972
Naive Bayes	0.9386796733212341	0.03174177045762223

Table 7: Tabela de resultados da primeira parte dos experimentos para o dataset Breast Cancer

Nela podemos ver que houve uma variação grande entre as médias dos classificadores mais simples, zeroR, oneR, oneR probabilístico, e os classificadores

mais complexos, o centroide e o naive bayes, sendo essa diferença maior que 0.3 entre o maior dos mais simples e o menor dos mais complexos, o único que foge disso foi o oneR centroide que mesmo sendo mais complexo teve a média na mesma faixa que o zeroR, além disso vemos que o naive bayes foi o melhor novamente, com uma média de 0.93 e com o menor desvio padrão dentre os mais complexos.

A figura 7 mostra o boxplot dos classificadores testados na primeira parte dos experimentos para a base de dados bresar cancer.

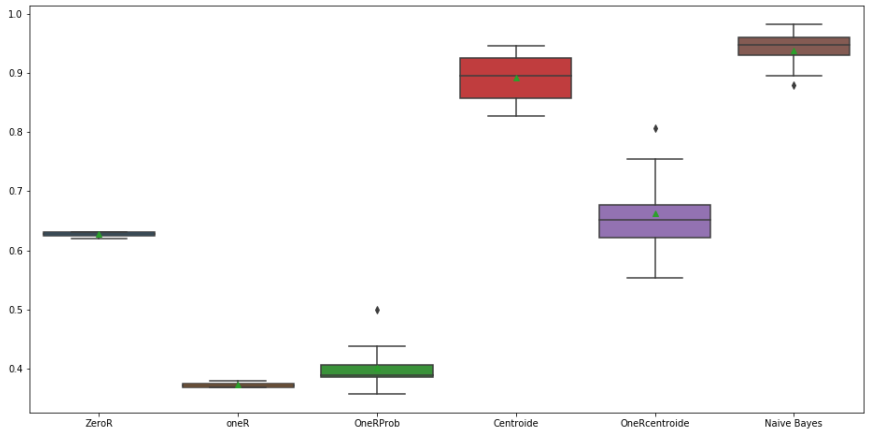

Figure 7: Boxplot da primeira parte dos experimentos para o dataset Breast Cancer

Nela vemos que o naive bayes foi o melhor em média, mas o centroide teve algumas acurácias tão boas quanto o naive bayes, além disso o naive teve um outlier que foi tão ruim quanto a média do centroide, mas por sua média ser a melhor e sua variância ser tão baixa ainda assim consideramos ele como sendo o melhor dentre os classificadores testados na primeira parte.

Agora para a segunda parte dos experimentos para a base de dados breast cancer, a tabela 8 mostra as médias e desvios padrões das acurácias, e os hiper-parâmetros selecionados dos classificadores em 10 folds do ciclo externo de treino e teste e 4 folds do ciclo interno de treino e validação.

Algoritmos	Média	Desvio padrão	Hiper-parâmetro
KNN	0.9316588886008124	0.03293529888963791	10
Arvore de decisão	0.910512272059459	0.04448546818983055	3
Rede neural	0.7918416731483882	0.201861685368878	15,200
Floresta de arvores	0.9633058940454584	0.03685810932160996	20

Table 8: Tabela de resultados da segunda parte dos experimentos para a base de dados Breast Cancer

Nela vemos que para essa base de dados os classificadores ficaram com médias próximas, menos a rede neural que teve o pior desempenho novamente, mas os outros tiveram todos médias acima de 0.9, mas o classificador floresta se destacou novamente com a melhor média e com o segundo melhor desvio padrão.

A figura 8 mostra o boxplot dos classificadores testados para a base de dados breast cancer

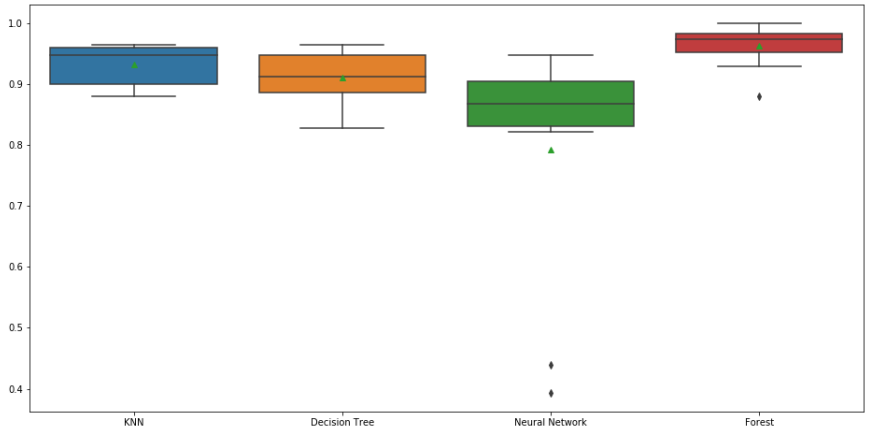

Figure 8: Boxplot da segunda parte do experimento para o dataset Breast Cancer

Nela vemos que o que fez com que a rede neural tivesse resultado pior foram dois outiliers ruins que trouxeram a média para baixo o suficiente para que ela fosse consideravelmente menor que a dos outros, além disso podemos ver que

o knn e o arvore de decisão foram quase iguais, mas o knn teve uma moda mais alta e então sua média foi um pouco maior que a da arvore de decisão, e vemos também que o floresta teve variância pequena na maior parte de seus folds mas teve um outlier ruim o suficiente para fazer com que sua média e variância ficassem similares a dos outros classificadores. Com isso concluímos que o melhor classificador para esta base de dados foi o floresta por sua média maior e desvio padrão baixo, mesmo com o outlier ruim

5. Conclusões

Neste trabalho vimos que houve uma discrepância grande entre os resultados dos algoritmos da primeira parte dos experimentos e da segunda parte, salvo o naive bayes, em geral os classificadores da primeira parte foram piores que os da segunda parte, enquanto os da primeira parte obtiveram médias inferiores a 0.70 na maioria das bases de dados enquanto os da segunda parte obtiveram resultados melhores que 0.8 na média para a maior parte das bases de dados, sendo a única base que eles tiveram resultados realmente comparaveis a base iris que foi a base mais simples usada nos testes, com menos atributos, menos classes e instancias, isso mostra a dificuldade do problema de classificação pois tentativas ingenuas de resolve-lo não tem bons resultados mesmo em base de dados não tão complexas como nós vemos no mundo real, sendo assim estas abordagens ingenuas devem ser apenas utilizadas para bases simples ou como benchmark para as técnicas mais complexas. Mas o classificador naive bayes teve bons resultados em todas as bases de dados mesmo sendo da primeira parte, isso vem do fato de embora naive estar no nome dele ele não seja tão ingenuo quanto os classificadores zeroR e oneR, utilizando métodos estatísticos para determinar a classe dos dados em vez de adivinhar essa classe baseado apenas na frequência de certos atributos dos dados como os classificadores zeroR e oneR fazem, o classificador centroide também obteve bons resultados, o que é de se esperar visto a semelhança dele com o classificador knn.

Também tivemos um resultado considerado surpreendente que foi a rede neu-

ral ter sido a pior na maior parte dos testes, mesmo ela sendo a mais usada no mercado e na academia hoje em dia, mas pensamos que isso aconteceu por causa dos hiper-parâmetros escolhidos visto que esse classificador é muito sensível a escolha de hiper-parâmetros. Outro resultado interessante foi o do oneR centroide que foi bom para algumas bases de dados e péssimo para outras, parece que colocar mesclar o oneR com o centroide fez com que ele fosse ligeiramente melhor em algumas bases, mas como a digitis mostrou isso pode não é regra pois ele teve resultado muito pior que o do centroide comum nessa base.

Ao final percebemos que os os classificador floresta foi o melhor de todos os classificadores testados nesta base de dados, sendo ele o que foi mais estavel, sempre mantendo média de acurácia acima de 0.9 e com pouca variancia em seus resultados.

Mas como falado anteriormente a rede neural foi a pior por causa da escolha de hiper-parâmetros, pode ser que com hiper-parâmetros diferentes ela tenha resultados melhores

Referencias bibliograficas

- Slides e Notas de Aula da Disciplina Disponibilizados pelo Professor por meio do *e-mail*.
- https://archive.ics.uci.edu/ml/datasets/iris
- https://archive.ics.uci.edu/ml/datasets/Pen-Based+Recognition+of+Handwritten+Digits
- https://archive.ics.uci.edu/ml/datasets/Wine
- https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(original)