Daniel Pereira Cinalli

Estudo de métodos de ensembles com seleção de classificadores dinâmica

Projeto de Graduação em Computação submetido à Universidade Federal do ABC para a obtenção dos créditos na disciplina Projeto de Graduação em Computação I do curso de Ciência da Computação

Orientador: Prof. Dr. Thiago Ferreira Covões

Universidade Federal do ABC

30 de Junho de 2021

RESUMO

Ensembles são utilizados pela sua maior capacidade de generalização e também melhor acurácia, quando comparados à utilização de um classificador apenas para a mesma tarefa. Uma abordagem possível é a seleção dinâmica dos classificadores de um ensemble, em que é estimado quais classificadores possuem um melhor desempenho, que então são utilizados para realizar a classificação.

Dois métodos de seleção dinâmica serão estudados. O primeiro escolhe os classificadores baseado na acurácia ao classificar objetos conhecidos próximos do objeto que se deseja classificar. O segundo utiliza o conceito de similaridade de decisão, escolhendo uma porcentagem dos classificadores que mais concordam na classificação de uma quantidade escolhida de objetos, gerados aleatoriamente na proximidade do objeto que se quer classificar.

Utilizando o teste de Friedman e consequentemente o teste de Nemenyi foi encontrado que nenhum dos métodos melhorou significativamente o desempenho do *ensemble*. Também se concluiu que os métodos não são significativamente diferentes entre si quanto ao seu desempenho.

SUMÁRIO

CAPÍTULO 1

INTRODUÇÃO

Em aprendizado de máquina, modelos preditivos são gerados à partir de um conjunto de dados, para que se possa fazer predições. Um modelo pode ser um classificador, caso se deseje realizar uma predição categórica, ou um regressor, quando o que se quer é uma predição numérica [Zhou 2012]. O foco deste estudo será em classificação, que de acordo com [Tan, Steinback e Kumar 2006], é a tarefa de dado um conjunto de entradas $\mathbf{X}$ e um conjunto de classes predefinidas $\mathbf{Y}$, de se aprender a função f que realiza o mapeamento $f : \mathbf{X} \rightarrow \mathbf{Y}$.

O desempenho de generalização de um classificador é definido como seu desempenho em classificar objetos não utilizados durante a fase de treinamento. Uma motivação do uso de *ensembles* é sua maior capacidade de generalização, se comparado à utilização de um único classificador para a mesma tarefa. Para obter isto é feita a combinação das predições dos diversos classificadores do *ensemble*, o que reduz o efeito daqueles classificadores que apresentam um desempenho insatisfatório [Polikar 2006].

O uso de *ensembles* visa tipicamente a melhora da acurácia em diversas tarefas. Em [Chowdhury et al. 2017] *ensembles* foram utilizadas para melhorar o desempenho na classificação de atividades físicas à partir de dados de acelerometria de pulso, quando comparados com classificadores individuais. Em [Yang et al. 2014] foi obtida uma predição de genes causadores de doenças mais acuradas com o uso de *ensembles*. Também foram utilizados para

a análise de imagens citológicas que podem ser utilizadas para a detecção precoce de câncer em [Filipczuk, Krawczyk e Woniak 2013], obtendo desempenho melhor do que o estado da arte alcançava no momento.

Entre os métodos de combinação para predições numéricas, pode-se citar os métodos de média simples e média ponderada. Para predições nominais, alguns métodos são voto da maioria, em que é necessário que pelo menos um dos resultados possíveis tenha recebido mais de metade dos votos; voto plural, onde simplesmente é escolhido aquele com mais votos; votação ponderada, entre outros. Existem, no entanto, diversos outros modos de se combinar os resultados de um *ensemble* [Zhou 2012].

Dado que um *ensemble* considera a predição de diversos classificadores, é desejável que este conjunto de classificadores possua alta diversidade. Diversidade neste contexto é difícil de se definir. Intuitivamente caso os classificadores individuais sejam similares, não haverá um ganho grande ao se combinar seus resultados. Ou seja, pode se entender por diversidade que os classificadores individuais que fazem parte do *ensemble* devem ter uma baixa correlação entre si. Classificadores com alta correlação entre si em geral diminuem o erro do *ensemble* menos se comparado a um *ensemble* onde estes possuem baixa correlação entre si, e se reduz ainda mais no caso de uma correlação negativa [Kuncheva e Whitaker 2003].

Uma alta diversidade portanto deve levar a uma melhor acurácia do *ensemble*, mas mensurar esta diversidade é difícil, e diversos métodos existentes não chegam a uma medida de diversidade que apresente boa correlação com a acurácia do *ensemble*. É entendido que esses métodos não sejam bons indicadores de diversidade, mas que a busca por diversidade ainda é importante [Zhou 2012].

Em classificação, existem também técnicas de seleção dinâmica de classificadores. Em seleção dinâmica, é feita uma estimação daqueles classificadores com melhor desempenho, que então são utilizados para realizar a classificação [Cruz, Sabourin e Cavalcanti 2018]. Neste trabalho, será estudado o desempenho de dois métodos onde os classificadores de um ensemble são escolhidos dinamicamente, que serão introduzidos a seguir.

O primeiro método escolhe os melhores classificadores de acordo com a acurácia local, utilizando os objetos de seu conjunto de treinamento e validação. Para isso, se escolhe os

objetos mais próximos do objeto que se deseja classificar, e se determina a acurácia com base nestes. À partir desses resultados, se determina quais classificadores serão utilizados para fazer a classificação.

O segundo método será baseado no artigo [Kurvers et al. 2019], onde foi analisada a ideia de se utilizar a similaridade de decisão para se encontrar indivíduos ou grupos com alto desempenho. Neste, a similaridade de decisão entre diferentes indivíduos foi usada para predizer a acurácia destes. Em [Cruz, Sabourin e Cavalcanti 2018], é proposto que enquanto é necessário ter diversidade entre os classificadores disponíveis em todo o ensemble, se deve ter consenso, e não diversidade, entre os classificadores selecionados para a tarefa de predição.

Embora estes métodos estejam escolhendo apenas alguns classificadores para realizar a predição, o objetivo é identificar bons classificadores localmente, e ainda é importante que o *ensemble* como um todo possua alta diversidade.

CAPÍTULO 2

JUSTIFICATIVA

O presente trabalho busca explorar duas técnicas em que se realiza a seleção dinâmica de classificadores em um ensemble. Essa técnica de seleção dinâmica tem como objetivo aumentar a acurácia de uma predição, selecionando por algum critério quais classificadores do ensemble serão utilizados para a predição sendo realizada para um dado objeto. Deseja-se determinar se as técnicas estudadas são viáveis como uma ferramenta para se obter uma melhor acurácia para um ensemble.

A técnica que utiliza o critério de similaridade de decisão foi baseada em [Kurvers et al. 2019]. No artigo se mostra que este critério consegue obter uma maior acurácia selecionando um subconjunto de todos os indivíduos de um grupo para se fazer uma predição, se comparado a utilizar o conhecimento combinado de todos. Dado o resultado positivo apresentado no artigo, a técnica será implementada em um ensemble para determinar seu desempenho no contexto de aprendizado de máquina.

Esses métodos vão contra a ideia de realizar uma predição com um conjunto de classificadores o mais diverso possível. No entanto como é proposto em [Cruz, Sabourin e Cavalcanti 2018], durante a seleção faz mais sentido se procurar consenso entre os classificadores. Intuitivamente pode-se pensar também que caso um subconjunto do *ensemble* tenha uma

acurácia maior em uma certa região do espaço de atributos, que estes classificadores possam levar a uma predição melhor quando comparada ao se utilizar todo o *ensemble*.

CAPÍTULO 3

FUNDAMENTAÇÃO TEÓRICA

Neste capítulo são apresentados os fundamentos de *ensembles*, os classificadores-base que serão utilizados, e os conceitos necessários para a análise dos dados.

3.1 Árvore de decisão

Uma árvore de decisão é composta de nós de decisão e de nós folha. Em cada nó de decisão, é implementada uma função de teste com saídas discretas, em que o resultado do teste determina qual o próximo nó escolhido. Isso ocorre até se encontrar um nó folha, que é um nó em que se encontra uma das possíveis saídas. Para se gerar uma árvore de decisão para classificação, se quantiza o quão boa é uma divisão criada por um nó de decisão usando uma medida de impureza. Uma medida possível é a função de entropia, que para um nó m e exemplo i é dada por [Alpaydin 2010]:

$$I_m = -\sum p_m^i log_2 p_m^i \tag{3.1}$$

Uma divisão pura, por exemplo, é uma em que a distribuição das classes após a divisão será (0, 1). A construção da árvore é feita se escolhendo a decisão com entropia mínima. Isso é feito recursivamente até que todas as divisões sejam puras. O tamanho do espaço de busca

faz com que seja inviável de se encontrar uma árvore ótima, e por isso algoritmos de construção de árvores de decisão usualmente utilizam uma estratégia gulosa que faz decisões localmente ótimas [Tan, Steinback e Kumar 2006].

3.2 Perceptron

Um perceptron possui n entradas e 1 saída. Para cada entrada x_j, está associado um peso w_j, e também um termo de *bias* w_0. O caso mais simples para se calcular a saída é a média ponderada, apresentada abaixo:

$$y = \sum_{j=1}^{n} w_j x_j + w_0 \tag{3.2}$$

O termo w_0 pode ser visto como se fosse associado a um termo x_0, com valor sempre igual a 1. A saída do perceptron pode então ser representada pelo produto escalar abaixo:

$$y = \mathbf{w}^T \mathbf{x} \tag{3.3}$$

em que $\mathbf{x} = [1, x_1, ..., x_n]$ e $\mathbf{w} = [1, w_1, ..., w_n]$. Durante o treinamento, se aprendem os pesos $\mathbf{w}$. Em treinamento online inicia-se com pesos aleatórios, a saída y é calculada para a primeira instância da entrada, os pesos são atualizados e o processo é repetido até que toda instância tenha passado pelo perceptron. Abaixo é mostrado o valor da atualização dos pesos, onde α é a taxa de aprendizado, $y^{(t)\prime}$ é a predição, $y^{(t)}$ é o valor real, e $x_j^{(t)}$ é a entrada, para a instância t:

$$\Delta w_j^{(t)} = \alpha(y^{(t)} - y^{(t)\prime}) x_j^{(t)} \tag{3.4}$$

3.3 Classificador Naïve Bayes

O teorema de Bayes é utilizado para se determinar a probabilidade de um evento Y ocorrer dado que se conhece que o evento X já ocorreu, e é dado pela seguinte equação:

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)} \tag{3.5}$$

Para se utilizar o teorema de Bayes para classificação, é necessário aprender as probabilidades a posteriori $P(Y|\mathbf{X})$, em que Y é a classe e $\mathbf{X}$ é o conjunto de atributos, à partir dos dados de treinamento. Sabendo essas probabilidades e sendo $\mathbf{X}'$ o objeto que quer se classificar, se encontra a classe Y' que maximiza a probabilidade a posteriori $P(Y'|\mathbf{X}')$. O teorema de Bayes é utilizado então para se expressar esta probabilidade em função de $P(Y)$ (probabilidade a priori), $P(\mathbf{X}|Y)$ (probabilidade condicional de classe), e $P(X)$ (evidência) [Tan, Steinback e Kumar 2006].

Em uma classificador Naïve Bayes, se supõe que os atributos são condicionalmente independentes. Com essa suposição se tem a seguinte relação:

$$P(\mathbf{X}|Y = y) = \prod_{i=1}^{n} P(X_i|Y = y) \tag{3.6}$$

Com essa suposição, não é necessário computar a probabilidade condicional de classe para cada combinação de $\mathbf{X}$, apenas computando a probabilidade condicional para cada X_i, dado um Y. Com isso, para se classificar um objeto, se usa a seguinte equação para se obter a probabilidade a posteriori de cada classe:

$$P(Y|\mathbf{X}) = \frac{P(Y)\prod_{i=1}^{n} P(X_i|Y)}{P(\mathbf{X})} \tag{3.7}$$

É suficiente se maximizar o numerador para se encontrar a classe do objeto, já que $P(\mathbf{X})$ será igual para todo Y.

3.4 Classificador kNN

O classificador kNN (*k-nearest neighbours*) computa a distância entre o objeto $\mathbf{X}$ que se deseja classificar e cada exemplo do conjunto de treinamento, e então seleciona os k mais próximos dentre estes, gerando o conjunto denotado por $D_{\mathbf{X}}$. O processo para se escolher a classe consiste então de uma votação majoritária, onde a classe em maior número em $D_{\mathbf{X}}$ é a predição para o objeto $\mathbf{X}$ [Tan, Steinback e Kumar 2006].

3.5 Ensembles

Classificadores distintos podem ser utilizados em conjunto para classificação no que é chamado um *ensemble*. *Ensembles* consistem de um conjunto de classificadores base treinados no conjunto de treinamento. A classificação realizada pelo *ensemble* é realizada por votação das predições feitas pelos classificadores base. Esse método tem como objetivo aumentar a acurácia de classificação, se comparado ao utilizar um de seus classificadores base apenas. Para se conseguir isso, há duas condições necessárias:

- Os classificadores base devem ser independentes entre si;
- Os classificadores base devem ter desempenho melhor que uma escolha aleatória.

Na primeira condição, dois classificadores serem independentes significa que seus erros não são correlacionados entre si. Na segunda condição, se implica que cada classificador precisa ter uma taxa de erros menor que 50%, no caso de classificação binária, com apenas duas classes e com classes balanceadas [Tan, Steinback e Kumar 2006]. No caso geral, para se obter um desempenho melhor que uma escolha aleatória a taxa de erro ϵ deve obedecer a

$$\epsilon < 1 - \frac{1}{N_c} \tag{3.8}$$

em que N_c é o número de classes, supondo que as classes são balanceadas. Por exemplo, para 10 classes a taxa de erro ao se escolher aleatoriamente é de 90%.

3.6 Bagging

Há diversas formas de se criar *ensembles*. Dentre estas pode-se alterar o algoritmo de treinamento da cada classificador-base, por meio de seus parâmetros. Outra possibilidade é a de se manipular o conjunto de treinamento, treinando cada classificador com um subconjunto amostrado do conjunto original. Um destes métodos é o *bagging* [Tan, Steinback e Kumar 2006].

O termo *bagging* se refere a *bootstrap aggregating*. Dado um *dataset* com m exemplos de treinamento, se toma deste uma amostra de m exemplos usando amostragem com substituição. Com essa abordagem, não se diminui a quantidade de exemplos que cada classificador do *ensemble* usará para ser treinado, evitando treinamento sobre amostras pouco representativas. Cada uma dessas amostras possui em média 63,2% do conjunto de treinamento original [Zhou 2012].

3.7 Validação cruzada

É necessário realizar a avaliação do desempenho de um classificador. Enquanto é possível simplesmente se definir um conjunto de treinamento e um de teste para com este se obter uma medida de desempenho (por exemplo acurácia), outros métodos possibilitam uma avaliação melhor. Um método possível é a validação cruzada.

Em validação cruzada, um conjunto de dados é particionado em k partes de mesmo tamanho. Uma dessas partições será o conjunto de teste enquanto as outras $k-1$ partes serão usadas para treinamento. Isso é repetido até todas as k partições terem sido utilizadas como conjunto de teste [Tan, Steinback e Kumar 2006].

3.8 Rank

Para realizar a análise e comparação de uma grande quantidade de resultados de desempenho de diversos classificadores, se torna necessário o uso de métodos estatísticos. Um método útil para esse cenário é o método de Friedman, que utiliza o *rank* médio de cada classificador.

Para cada par de classificador e *dataset*, se tem uma medida de desempenho obtida no experimento, como por exemplo a acurácia. Para um *dataset*, um *rank* é dado para cada classificador, sendo *rank* 1 para a maior acurácia, *rank* 2 para a segunda maior, e assim por diante. Em caso de um empate, a média entre os *ranks* é atribuída. Por exemplo, caso os 2 melhores classificadores empatem, seus *ranks* serão 1,5, a média entre 1 e 2. Ao final, se

obtém o *rank* médio para um classificador pela média de seus *ranks* sobre todos os *datasets*. Para cada classificador seu *rank* médio é dado por

$$R_j = \frac{1}{N}\sum_i r_i^j \tag{3.9}$$

em que r_i^j o *rank* do j-ésimo classificador para o i-ésimo *dataset*.

3.9 Teste de Friedman

Em estatística, uma hipótese nula é uma hipótese que se procura rejeitar. O teste de Friedman utiliza a hipótese nula de que todos os classificadores são equivalentes, ou seja, os *ranks* médios de todos os classificadores são iguais. Sendo N e k respectivamente a quantidade de classificadores e de *datasets*, caso a hipótese nula seja verdadeira a estatística de Friedman

$$\chi_F^2 = \frac{12N}{k(k+1)}\left[\sum_j R_j^2 - \frac{k(k+1)^2}{4}\right] \tag{3.10}$$

será distribuída de acordo com χ_F^2 com $k-1$ graus de liberdade, para N e k suficientemente grandes. Sendo a hipótese nula do teste de Friedman rejeitada, se conclui que existe diferenças significativas entre classificadores analisados. Com isso, para se determinar as diferenças par a par se seguem com outros testes estatísticos.

3.10 Teste de Nemenyi

O teste de Nemenyi é utilizado após o teste de Friedman. No teste de Nemenyi dois classificadores tem um desempenho significantemente diferente caso seus *ranks* tenham entre si uma diferença maior que a diferença crítica CD, dada pela equação

$$CD = q_\alpha\sqrt{\frac{k(k+1)}{6N}} \tag{3.11}$$

em que q_α é dado pela tabela de amplitude estudentizada dividida por $\sqrt{2}$. N se refere à quantidade de *datasets* e k à quantidade de classificadores.

Os resultados do teste de Nemenyi são representados graficamente pelo diagrama de diferença crítica. No diagrama se mostra todos os *ranks* médios dos classificadores, com barras juntando os classificadores sem diferença significativa. Uma barra no topo representa o valor da diferença crítica CD [Demšar 2006].

CAPÍTULO 4

MÉTODOS DE SELEÇÃO DINÂMICA

Neste capítulo serão apresentados em maior detalhe como os métodos estudados funcionam. Ao final se apresentam como os experimentos serão realizados e como será feita a análise dos resultados

4.1 Método de seleção por acurácia

Este método escolhe os melhores classificadores de acordo com a acurácia local, com base nos objetos de seu conjunto de treinamento e validação. Possui dois parâmetros: N, o número de objetos próximos que serão considerados; e $Acc_{\min}$, a acurácia mínima necessária para se aceitar um classificador. Para se classificar um objeto $\mathbf{x}$ encontra-se os N pontos do conjunto de treinamento com menor distância euclidiana de $\mathbf{x}$ e se obtém a acurácia de cada classificador com base nestes. Os classificadores com maior acurácia que o mínimo são então escolhidos para classificarem $\mathbf{x}$ com uma média ponderada das classificações de cada classificador escolhido, e o resultado é normalizado. Como é possível nenhum classificador possuir a acurácia necessária, neste caso escolhe-se apenas aquele com maior acurácia. Abaixo se apresenta o pseudo-código do método.

Algorithm 1: Método de seleção por acurácia

```
Seja o objeto que se deseja classificar x.
Seja o conjunto de treinamento X com suas classes Y.
Seja o número de objetos próximos que serão usados N.
Seja o conjunto de classificadores já treinados C.
Seja a acurácia mínima Acc_min.
próximos = N elementos de X com menor distância euclidiana de x.
yPreds = {C_i.predição(próximos) para cada C_i em C}.
accs = calculaAcurácia(yPreds, Y)
clfsEscolhidos = {clf_i | acc_i > Acc_min}
accsEscolhidos = {acc_i | acc_i > Acc_min}
if clfsEscolhidos não estiver vazio then
    acuraciaTotal = soma(accs)
    pesos = {peso_i = acc_i / acuraciaTotal}
    predicãoFinal = predicãoPonderada(P, clfsEscolhidos, pesos)
else
    clfEscolhido = classificador com maior acurácia dentre C
    predicãoFinal = predição(x, clfEscolhido)
end
return predicãoFinal
```

4.2 Método de seleção por similaridade

Diferentemente do método anterior, que busca objetos próximos já existentes, os objetos serão gerados aleatoriamente, próximos ao objeto que se quer classificar, seguindo uma distribuição normal para cada atributo com média igual ao valor de cada atributo para **x** e com variância igual a 10% da variância original de cada atributo obtida do conjunto de treinamento, e a decisão de quais serão os classificadores escolhidos será baseado na similaridade de decisão. Os parâmetros para este método são: N, a quantidade de objetos aleatórios que serão gerados; M, a quantidade de classificadores que serão escolhidos.

A similaridade de decisão é medida apresentando um conjunto de perguntas que cada indivíduo deve responder. No contexto de classificação, essas perguntas são as classificações dos objetos gerados. Para cada indivíduo, se calcula a porcentagem de concordância em relação a cada outro indivíduo, ou seja, o quanto concordam nas respostas às perguntas. A similaridade de decisão de cada indivíduo será a média das porcentagens de concordância calculadas para este. Tanto na análise matemática quanto na empírica, se concluiu que a

abordagem é viável, encontrando grupos com melhor desempenho [Kurvers et al. 2019]. Abaixo está o pseudo-código deste método.

Algorithm 2: Método de seleção por similaridade

Seja o objeto que se deseja classificar **x**.
Seja V_i a variância de cada atributo i do conjunto de treinamento.
Seja N_a a quantidade de atributos.
Seja N o número de classificadores que serão escolhidos para realizar a classificação.
Seja M a quantidade de classificadores do *ensemble*.
Seja K a quantidade de objetos que serão gerados.
Seja C o conjunto de classificadores do *ensemble*.
$yPreds_i^k$ se refere à predição do classificador i do objeto k.
varsGerados = { $V_i \cdot 10\%$ | i=1, 2, ..., N_a}
objetosGerados = K objetos gerados utilizando uma distribuição normal para cada atributo de acordo com as variâncias em varsGerados com centro em **x**
similaridades = {$s_i = 0$ | i = 1, 2, ..., M}
for *i em 1, 2, ..., M* **do**
    $yPreds_i = C_i.predicao(objetosGerados)$
for *i em 1, 2, ..., M* **do**
    for *j em 1, 2, ..., M* **do**
        for *k em 1, 2, ..., K* **do**
            if $yPreds_i^k = yPreds_j^k$ **then**
                $similaridades_i + 1$
$C_{escolhidos}$ = N classificadores com maior similaridade de decisão de acordo com o vetor similaridades
predição Final = predição de **x** considerando os classificadores em $C_escolhidos$ usando votação
return predição Final

CAPÍTULO 5

EXPERIMENTOS E ANÁLISE

Os experimentos foram realizados sobre 10 *datasets*, utilizando 20 classificadores-base em cada *ensemble*, variando-se o parâmetro relevante de cada método. Em cada experimento é utilizado 5 *folds* para validação cruzada. São utilizados 4 tipos de classificadores-base: Árvores de decisão, perceptrons, Naïve Bayes, e kNN. Além desses, foi também utilizado um *ensemble* misto composto de classificadores de árvores de decisão, perceptrons e kNN, para se estudar a performance do método com um *ensemble* misto. O classificador Naïve-Bayes não foi inserido no *ensemble* misto por apresentar alto tempo de processamento e desempenho relativamente pior que os outros em testes preliminares, e não se achou necessário incluí-lo pois a quantidade de classificadores-base diferentes já era o suficiente.

Para se obter classificadores-base mais variados, se utilizou *bagging*. Além disso, se variou os parâmetros em sua criação. Para kNN, se variou o parâmetro k de 1 a 10, e se variou os pesos dados para cada vizinho entre uniforme e inversamente proporcional à distância, em ordem, gerando 20 combinações. Para o perceptron se utilizou regularização L2, variando o coeficiente de regularização entre 0 e 0,1 aleatoriamente em uma distribuição uniforme. Para a árvore de decisão se configurou de forma que a escolha da próxima divisão não escolha a melhor possível e sim que escolha aleatoriamente com probabilidades proporcionais ao quão boa é aquela divisão. Para o Naïve-Bayes se alterou a variável *var_smoothing*

que adiciona variância à variância original obtida pelos dados de treinamento, suavizando a curva. Esse parâmetro foi variado de forma aleatória uniforme entre $1 \cdot 10^{-10}$ $1 \cdot 10^{-8}$.

O *ensemble* misto é composto de um terço de árvores de decisão, um terço de perceptrons, arredondados para baixo e o restante classificadores kNN. Isso resulta em respectivamente 6, 6 e 8 classificadores.

Para o método de seleção por acurácia, se varia a acurácia mínima necessária de 0% a 100%, em passos de 5%, em que 0% é equivalente a se utilizar o *ensemble* completa. A quantidade de objetos próximos verificados não foi variada e foi de 10 objetos.

Para o método de seleção por similaridade, se varia a porcentagem de classificadores que serão escolhidos de 5% a 100%, em passos de 5%, em que 100% é equivalente a se utilizar o ensemble completo. Foram gerados 10 objetos próximos ao ponto que se deseja classificar em todos os casos.

Para a análise dos resultados foi realizado o teste de Friedman seguido do teste de Nemenyi quando apropriado para verificar se existem diferenças significativas, sendo que para visualização da análise foi utilizado o diagrama de diferença crítica. Também se utilizou o *rank* médio para comparação, representando graficamente por meio de gráficos de tendência entre *rank* obtido e parâmetro utilizado no método.

CAPÍTULO 6

RESULTADOS

Primeiramente, os resultados do método de seleção por acurácia foram analisados. Com as acurácias obtidas nos experimentos se aplicou o teste de Friedman. O *p-value* obtido foi de $5,59x10^{-182}$. A hipótese nula de que os classificadores utilizados são equivalentes é então provada errada. Pode-se então seguir com o teste de Nemenyi.

Com os *ranks* dos classificadores se elaborou o gráfico apresentado na Figura 6.1. Nota-se que *mix* se refere ao ensemble misto. Pode se observar uma tendência para *ranks* melhores conforme se diminui a acurácia mínima para os *ensembles* compostos de árvore de decisão e o misto. Isso implica que ao se diminuir a acurácia mínima, ou seja, filtrar menos classificadores, o desempenho melhora nestes casos. O *ensemble* composto de classificadores kNN apresenta uma pequena tendência positiva, enquanto o perceptron não parece apresentar uma tendência. Já Naïve-Bayes apresenta uma tendência negativa, indicando que ao se aumentar a acurácia mínima o desempenho tende a melhorar, mesmo que isso esteja potencialmente filtrando mais classificadores. No entanto o desempenho do Naïve-Bayes é ruim comparado aos outros. Nota-se que a escolha do classificador-base de cada *ensemble* apresentou uma influência maior no *rank* do que a escolha do parâmetro de acurácia mínima.

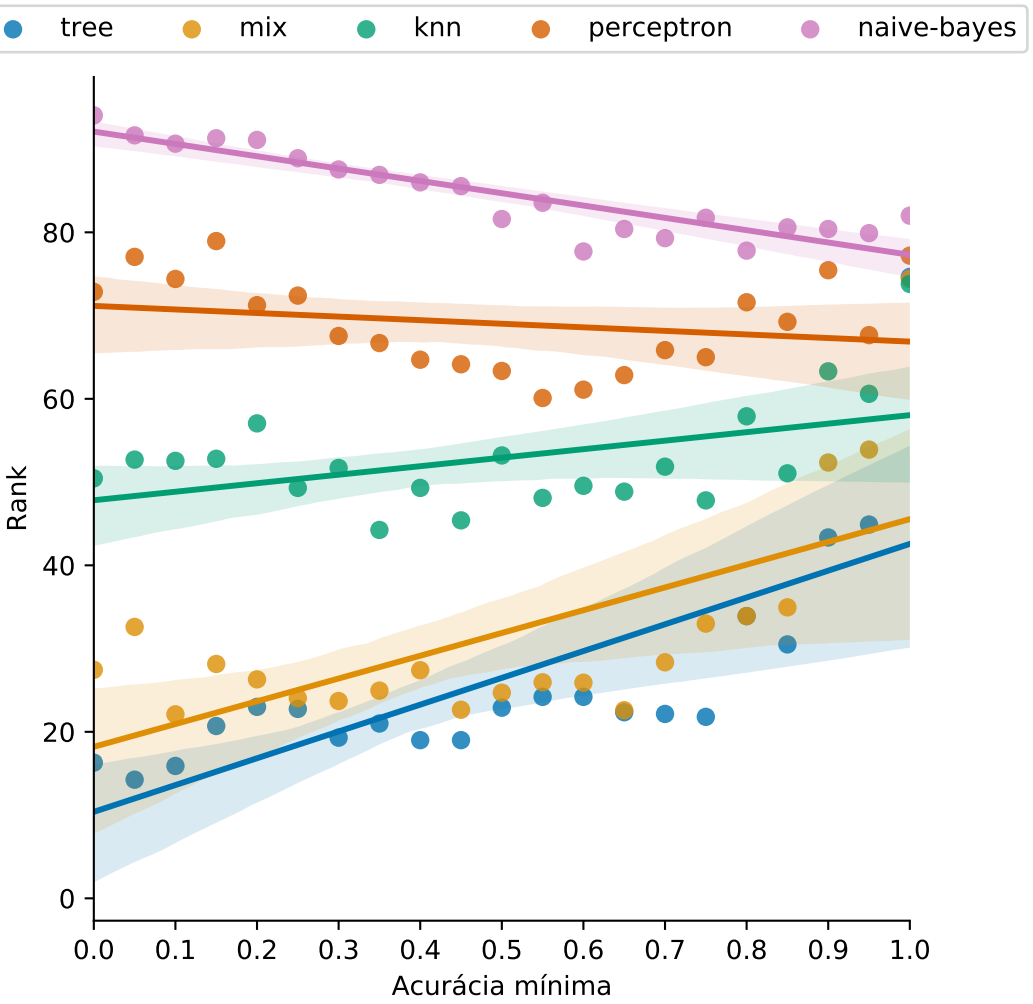

Figura 6.1: Gráfico de dispersão para o método da acurácia

Segue-se então com o teste de Nemenyi. O valor de α utilizado é de 0,05. Os resultados são mostrados no diagrama de diferença crítica na Figura 6.2. Como há muitos classificadores para se comparar em um diagrama só, o melhor e o pior *rank* para cada classificador-base foram selecionados, de acordo com a Tabela A.1. Na Figura 6.2, pode ser observado que apenas para o classificador-base árvore de decisão há uma diferença significativa entre seu pior e melhor caso. Para os outros, no entanto, se conclui que variar o parâmetro de acurácia mínima não obteve resultados significativos.

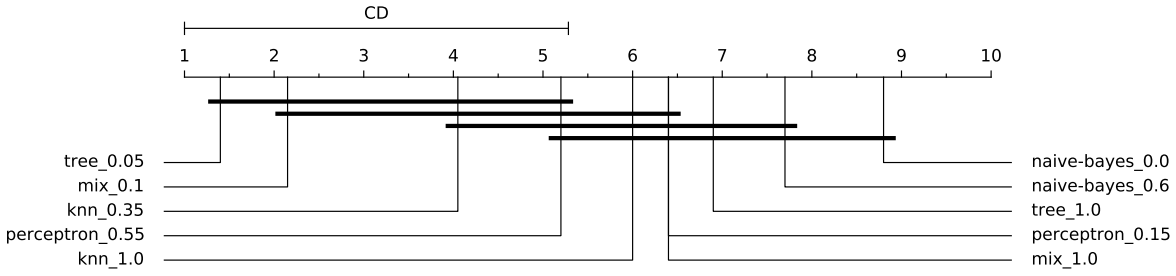

Figura 6.2: Diagrama de diferença crítica para o método da acurácia

A análise feita é repetida para o método de seleção por similaridade. O teste de Friedman resulta em um *p-value* de $1,12x10^{-169}$. Com isso se conclui que a hipótese nula está errada e que há diferenças significativas entre os métodos.

A Figura 6.3 apresenta as tendências do *rank* para o método de similaridade. Há uma tendência negativa para os *ensembles* com classificadores-base árvores de decisão, kNN e misto. Os outros *ensembles* não apresentam uma tendência. Novamente se percebe que a escolha de classificador-base apresenta maior influência do que o parâmetro.

Seguindo-se com o teste de Nemenyi, também com valor de α de 0,05 se conclui que a variação do parâmetro de classificadores escolhidos não obteve diferenças significantes. O diagrama de diferença crítica na Figura 6.4 demonstra estes resultados. Como no caso anterior há muitos classificadores e portanto apenas aqueles com o melhor e pior *rank* para cada classificador-base são mostrados. Na Tabela A.2 os *ranks* completos podem ser visualizados.

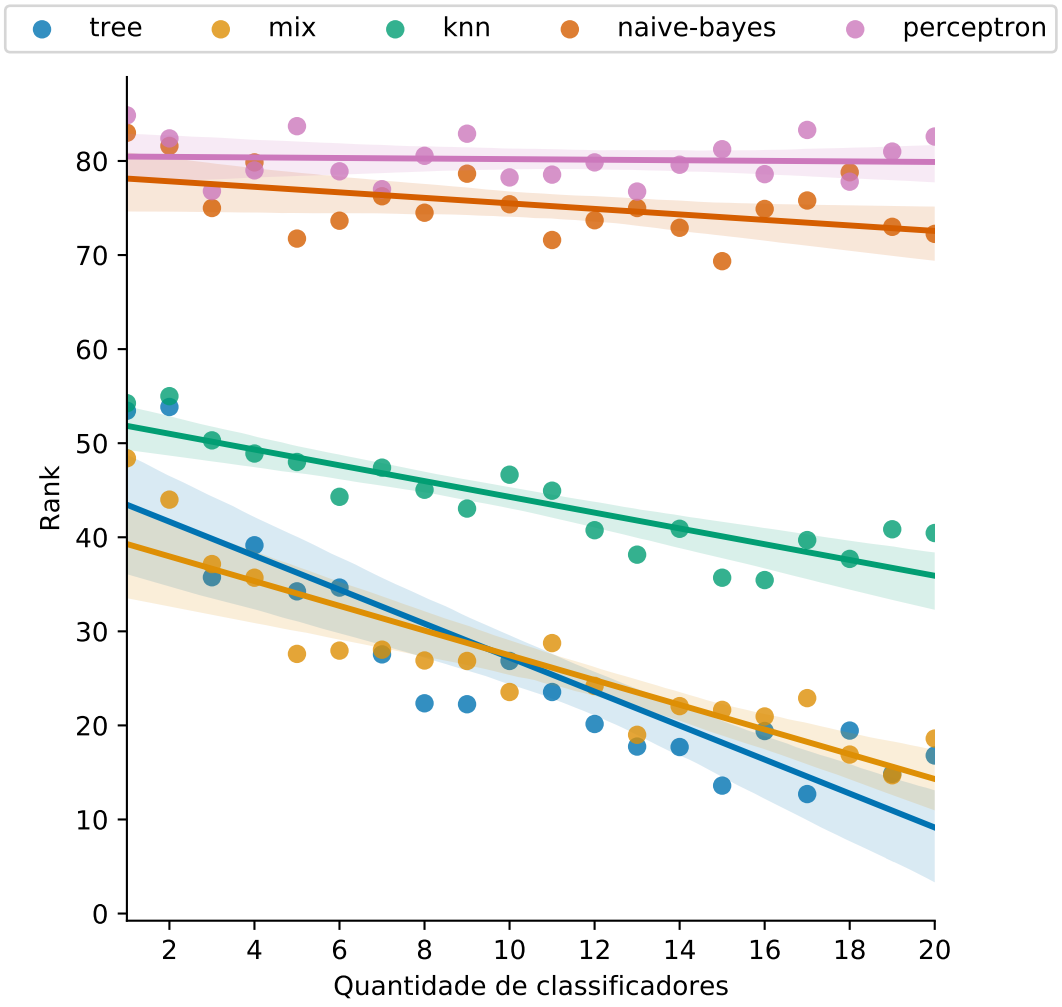

Figura 6.3: Gráfico de dispersão para o método da similaridade

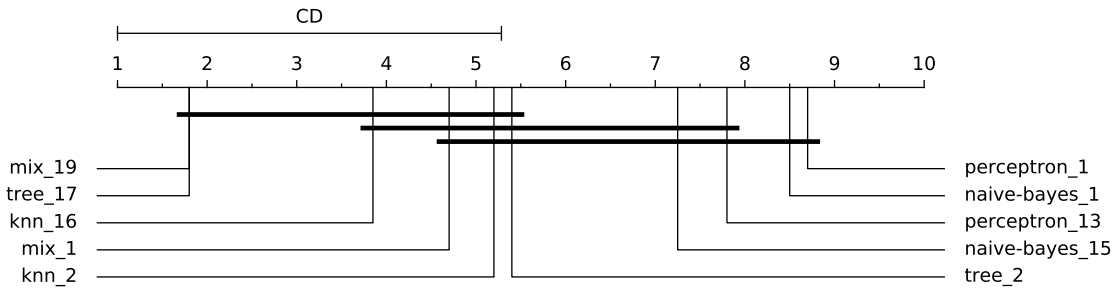

Figura 6.4: Diagrama de diferença crítica para o método da similaridade

Os resultados anteriores indicam que nenhum dos métodos melhoram o desempenho se comparados a utilizar o *ensemble* inteiro. No entanto ainda é possível se comparar os dois métodos entre si e na Figura 6.5 se visualiza o diagrama de diferença crítica entre os classificadores comparados anteriormente. Para os classificadores-base misto, kNN, perceptron e

Naïve-Bayes, não existem diferenças significativas entre os diferentes métodos e parâmetros utilizados. Para a árvore de decisão, se percebe que o método de acurácia com uma acurácia mínima de 100% não possui diferenças significativas com o método de similaridade com 2 classificadores, mas é significativamente diferente dos outros dois classificadores. Com base no gráfico se conclui que não há uma vantagem entre a utilização de um método sobre o outro.

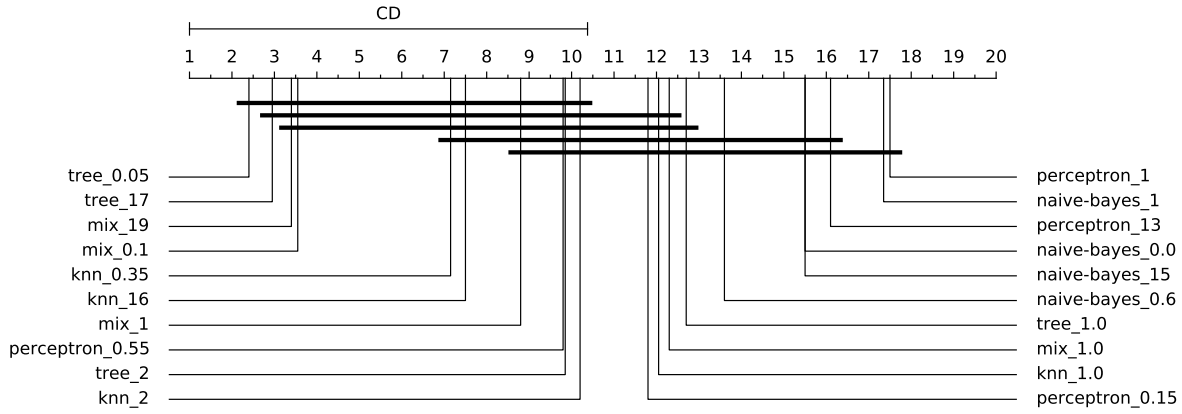

Figura 6.5: Diagrama de diferença crítica comparando os dois métodos

As próximas figuras são *boxplots* das acurácias obtidas para algumas das configurações de método estudado e classificador-base. As demais figuras se encontram no apêndice.

Na Figura 6.6, para o *dataset Electrical Grid Stability*, uma região entre 5 e 12 classificadores escolhidos, que apresenta uma maior acurácia se comparada ao se usar todos os classificadores. Para *HTRU Pulsar*, *Rice* e *GAMMA Telescope* se percebe um pequeno aumento na acurácia entre 2 e 3 classificadores selecionados. Os outros *datasets* não apresentam diferenças notáveis em relação à escolha de parâmetro.

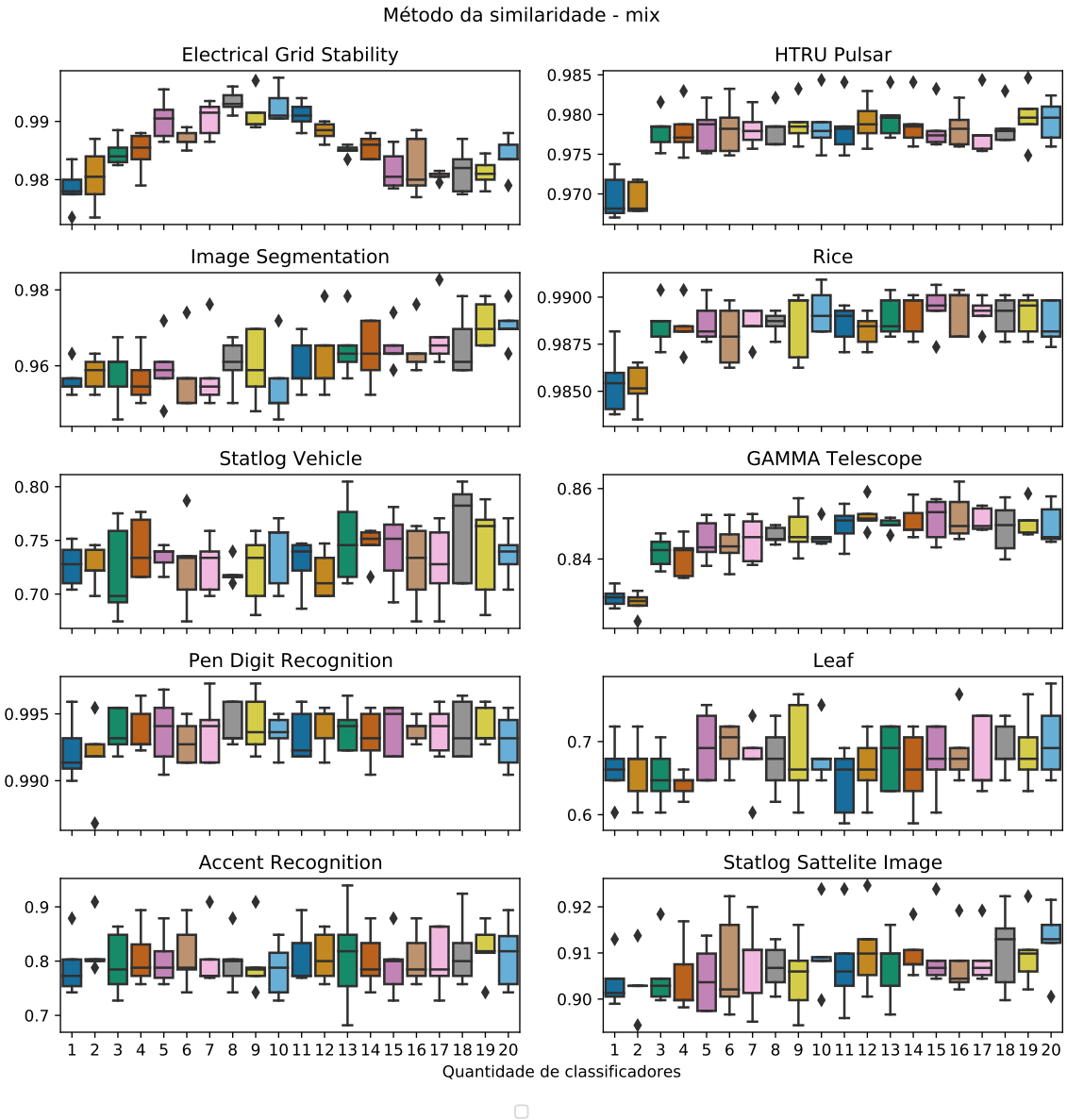

Figura 6.6: Resultados do método de similaridade com classificador-base misto

Na Figura 6.7, pode se observar que ao se selecionar uma acurácia mínima de 100%, há uma queda na acurácia. Isso se deve ao fato que é possível conforme se aumente a acurácia mínima, de que nenhum classificador-base consiga atingir esse mínimo. Neste caso, o algoritmo escolhe apenas o melhor classificador, para que seja possível continuar com a predição. Apesar desta queda brusca não se observa uma diminuição da acurácia gradual ao se aumentar a acurácia mínima em todos os casos. Mas é possível notar esta perda gradual de acurácia nos *datasets Accent Recognition, Leaf*, e levemente em *GAMMA Telescop* e *Statlog Sattelite Image*.

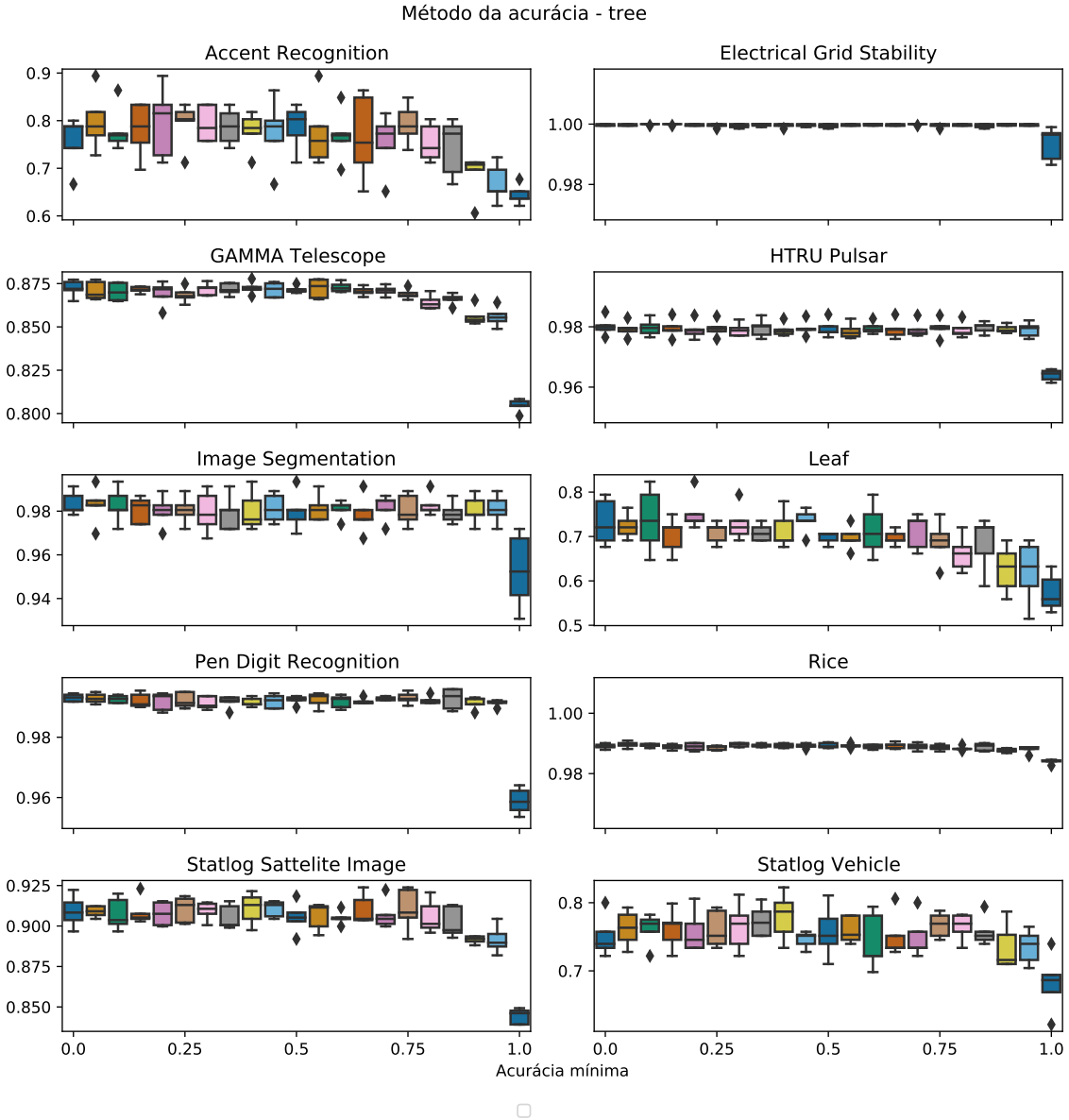

Figura 6.7: Resultados do método de acurácia com classificador-base árvore de decisão

Na Figura 6.8 se nota a tendência de se melhorar a acurácia conforme se aumenta o número de classificadores escolhidos nos *datasets Image Segmentation, Pen Digit Recognition, Rice, GAMMA Telescope* e *Rice*. Esse aumento parece ocorrer apenas no início, quando há poucos classificadores escolhidos. No entanto esse padrão não é observado em todos os *datasets*.

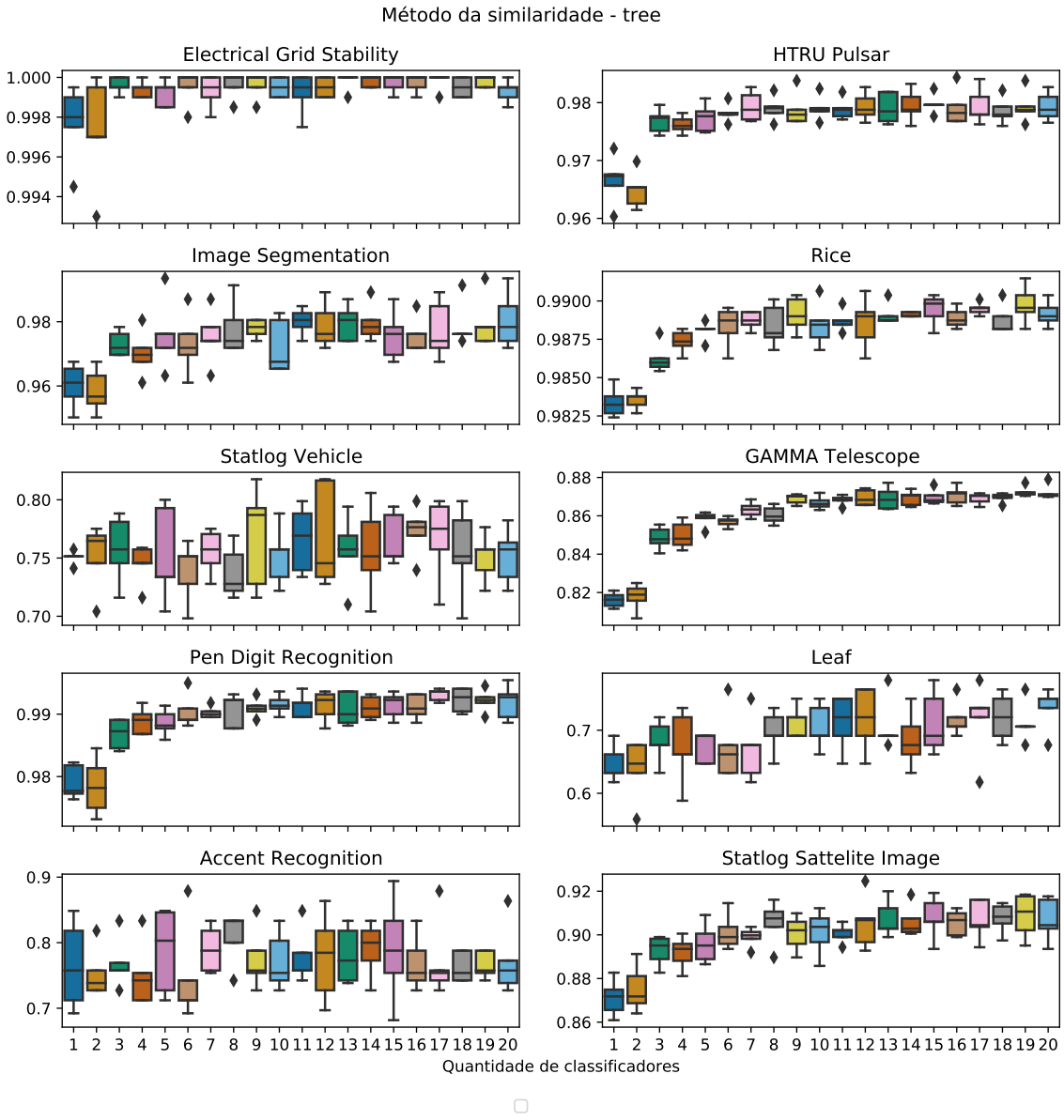

Figura 6.8: Resultados do método de similaridade com classificador-base árvore de decisão

CAPÍTULO 7

CONCLUSÃO

Dois métodos de seleção dinâmica em *ensembles* foram estudados. O método de seleção por acurácia que escolhe os classificadores de acordo com a acurácia de predição em pontos conhecidos do conjunto de treinamento; e o método de seleção por similaridade que seleciona os classificadores baseado na similaridade de decisão. A comparação entre estes métodos foi realizada utilizando técnicas estatísticas com o propósito de encontrar se diferenças entre os classificadores são significativas.

Com base nos resultados obtidos para o método de seleção por acurácia é possível concluir que dentre os classificadores estudados apenas Naïve-Bayes apresenta uma melhora ao se aumentar a acurácia mínima necessária para seleção. Para os outros classificadores, parece não se ter um ganho em desempenho em se utilizar o método. Apenas para o caso com o classificador-base árvore de decisão se obteve diferenças significativas ao se variar a acurácia mínima.

O método de seleção por similaridade não apresenta melhora ao se escolher uma quantidade menor de classificadores. Os resultados indicam que ao não se utilizar o método se tende a um melhor desempenho, embora a diferença ao se variar a quantidade de classificadores escolhidos não seja significativa. Entre os dois métodos estudados também não se encontrou diferenças significativas que indiquem um método como sendo o melhor.

Trabalhos futuros podem alterar os classificadores-base utilizados e verificar potenciais melhoras no desempenho. Pode-se também alterar os parâmetros dos métodos: para o método de seleção por acurácia a quantidade de pontos próximos utilizados para cálculo da acurácia e no método de seleção por similaridade pode se variar a quantidade de objetos gerados próximos ao objeto que se deseja classificar. Outra possibilidade é se encontrar outras medidas de similaridade e analisar se apresentam uma melhora no desempenho do método.

REFERÊNCIAS

[Alpaydin 2010]ALPAYDIN, E. *Introduction to Machine Learning*. [S.l.: s.n.], 2010.

[Chowdhury et al. 2017]CHOWDHURY, A. et al. Ensemble methods for classification of physical activities from wrist accelerometry. *Med Sci Sports*, p. 49, 2017.

[Cruz, Sabourin e Cavalcanti 2018]CRUZ, R. M. O.; SABOURIN, R.; CAVALCANTI, G. D. C. Dynamic classier selection: Recent advances and perspectives. *Information Fusion*, v. 41, p. 195–216, 2018.

[Demšar 2006]DEMŠAR, J. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, v. 7, n. 1, p. 1–30, 2006. Disponível em: <http://jmlr.org/papers/v7/demsar06a.html>.

[Filipczuk, Krawczyk e Woniak 2013]FILIPCZUK, P.; KRAWCZYK, B.; WONIAK, M. Classifier ensemble for an effective cytological image analysis. *Pattern Recognition Letters*, v. 34, n. 14, p. 1748–1757, 2013. ISSN 0167-8655. Innovative Knowledge Based Techniques in Pattern Recognition. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0167865513001888>.

[Kuncheva e Whitaker 2003]KUNCHEVA, L.; WHITAKER, C. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine Learning*, v. 51, p. 181207, 2003.

[Kurvers et al. 2019]KURVERS, R. H. J. M. et al. How to detect high-performing individuals and groups: Decision similarity predicts accuracy. *Science Advances*, American Association for the Advancement of Science, v. 5, n. 11, 2019. Disponível em: <https://advances.sciencemag.org/content/5/11/eaaw9011>.

[Polikar 2006]Polikar, R. Ensemble based systems in decision making. *IEEE Circuits and Systems Magazine*, v. 6, n. 3, p. 21–45, 2006.

[Tan, Steinback e Kumar 2006]TAN, P.-N.; STEINBACK, M.; KUMAR, V. *Introduction to Data Mining*. [S.l.: s.n.], 2006.

[Yang et al. 2014]YANG, P. et al. Ensemble positive unlabeled learning for disease gene identification. *PLoS ONE*, v. 5, 2014. Disponível em: <https://doi.org/10.1371/journal.pone.0097079>.

[Zhou 2012]ZHOU, Z.-H. *Ensemble methods: Foundations and algorithms*. [S.l.]: Taylor Francis Group, LLC, 2012.

APÊNDICE A

TABELAS DOS RANKS

Tabela A.1: Resultados do método de acurácia

kNN		Mix		Naïve-Bayes		Perceptron		Árvore de decisão	
Rank	Parâmetro	Rank	Parâmetro	Rank	Parâmetro	Rank	Parâmetro	Rank	Parâmetro
44,25	0,35	22,10	0,10	77,70	0,60	60,10	0,55	14,25	0,05
45,40	0,45	22,60	0,65	77,80	0,80	61,10	0,60	15,90	0,10
47,80	0,75	22,65	0,45	79,30	0,70	62,85	0,65	16,30	0,00
48,10	0,55	23,70	0,30	79,90	0,95	63,35	0,50	19,00	0,45
48,85	0,65	24,05	0,25	80,40	0,90	64,15	0,45	19,00	0,40
49,30	0,40	24,70	0,50	80,40	0,65	64,70	0,40	19,30	0,30
49,30	0,25	24,95	0,35	80,60	0,85	65,00	0,75	20,70	0,15
49,55	0,60	25,90	0,60	81,60	0,50	65,85	0,70	21,00	0,35
50,45	0,00	25,95	0,55	81,75	0,75	66,70	0,35	21,80	0,75
51,05	0,85	26,30	0,20	82,00	1,00	67,55	0,30	22,15	0,70
51,70	0,30	27,40	0,40	83,55	0,55	67,65	0,95	22,35	0,65
51,85	0,70	27,45	0,00	85,55	0,45	69,25	0,85	22,75	0,25
52,55	0,10	28,15	0,15	86,00	0,40	71,25	0,20	22,90	0,50
52,70	0,05	28,35	0,70	86,90	0,35	71,60	0,80	23,00	0,20
52,80	0,15	32,60	0,05	87,55	0,30	72,40	0,25	24,20	0,60
53,20	0,50	33,00	0,75	88,90	0,25	72,85	0,00	24,20	0,55
57,05	0,20	33,90	0,80	90,65	0,10	74,40	0,10	30,50	0,85
57,90	0,80	34,95	0,85	91,10	0,20	75,45	0,90	33,90	0,80
60,60	0,95	52,35	0,90	91,30	0,15	77,05	0,05	43,35	0,90
63,30	0,90	53,90	0,95	91,65	0,05	77,20	1,00	44,90	0,95
73,80	1,00	74,40	1,00	94,05	0,00	78,95	0,15	74,65	1,00

Tabela A.2: Resultados do método de similaridade

kNN		Mix		Naïve-Bayes		Perceptron		Árvore de decisão	
Rank	Parâmetro	Rank	Parâmetro	Rank	Parâmetro	Rank	Parâmetro	Rank	Parâmetro
35,45	16	14,70	19	69,35	15	76,75	13	12,70	17
35,70	15	16,90	18	71,60	11	76,80	3	13,60	15
37,70	18	18,60	20	71,75	5	77,00	7	14,85	19
38,15	13	19,00	13	72,25	20	77,80	18	16,80	20
39,70	17	20,95	16	72,90	14	78,25	10	17,70	14
40,45	20	21,65	15	73,00	19	78,55	11	17,75	13
40,75	12	22,05	14	73,65	6	78,60	16	19,40	16
40,85	19	22,90	17	73,70	12	78,90	6	19,45	18
40,90	14	23,55	10	74,50	8	79,00	4	20,15	12
43,05	9	24,20	12	74,90	16	79,60	14	22,25	9
44,30	6	26,85	9	75,00	13	79,85	12	22,35	8
44,95	11	26,90	8	75,00	3	80,55	8	23,55	11
45,05	8	27,60	5	75,40	10	81,00	19	26,85	10
46,65	10	27,95	6	75,80	17	81,25	15	27,55	7
47,40	7	28,05	7	76,25	7	82,40	2	34,25	5
48,00	5	28,75	11	78,65	9	82,60	20	34,65	6
48,90	4	35,70	4	78,80	18	82,90	9	35,75	3
50,30	3	37,15	3	79,85	4	83,30	17	39,15	4
54,25	1	44,00	2	81,60	2	83,70	5	53,45	1
55,00	2	48,40	1	83,00	1	84,85	1	53,85	2

APÊNDICE B

BOXPLOTS DOS RESULTADOS

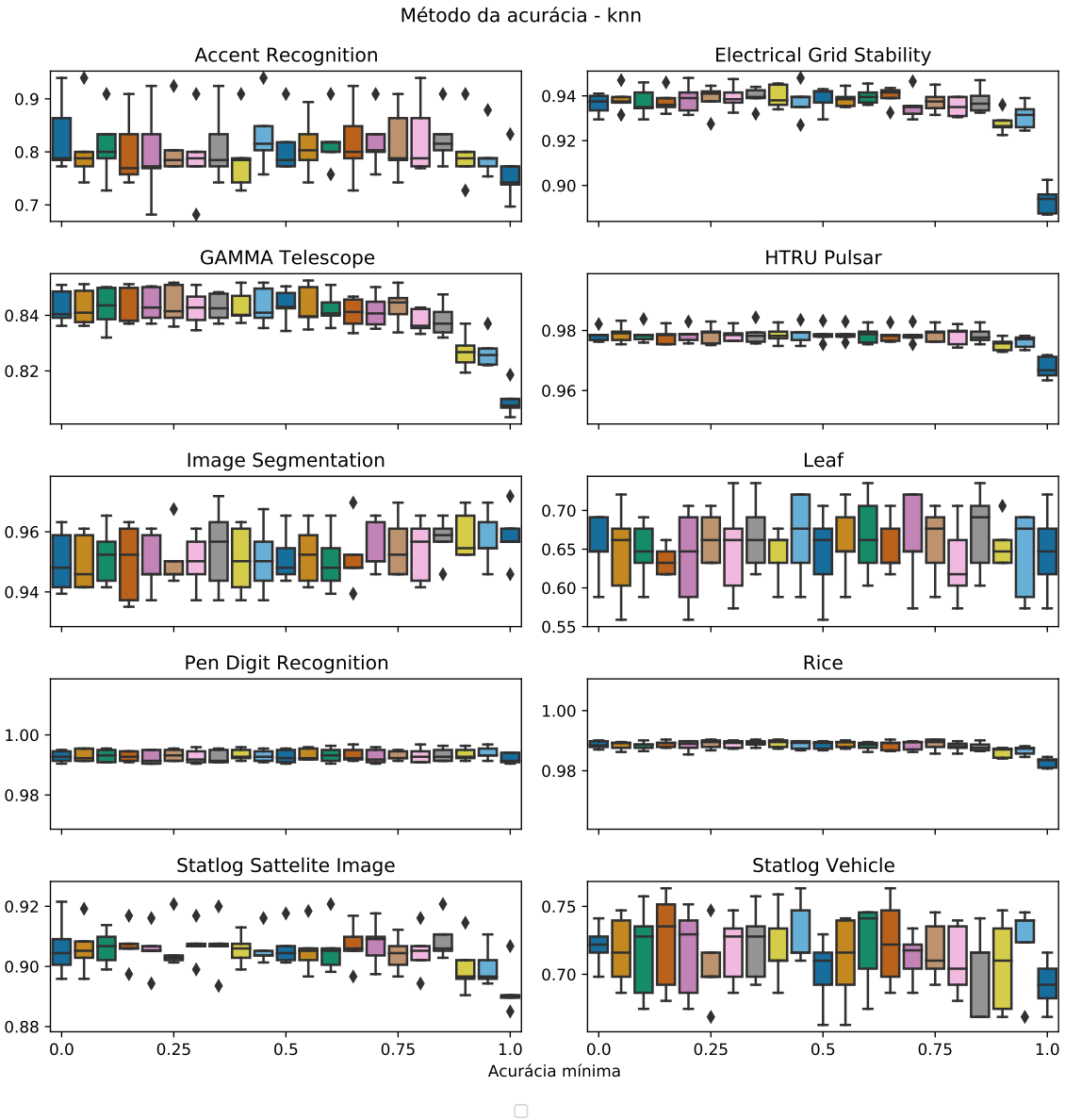

Figura B.1: Resultados do método de acurácia com classificador-base kNN

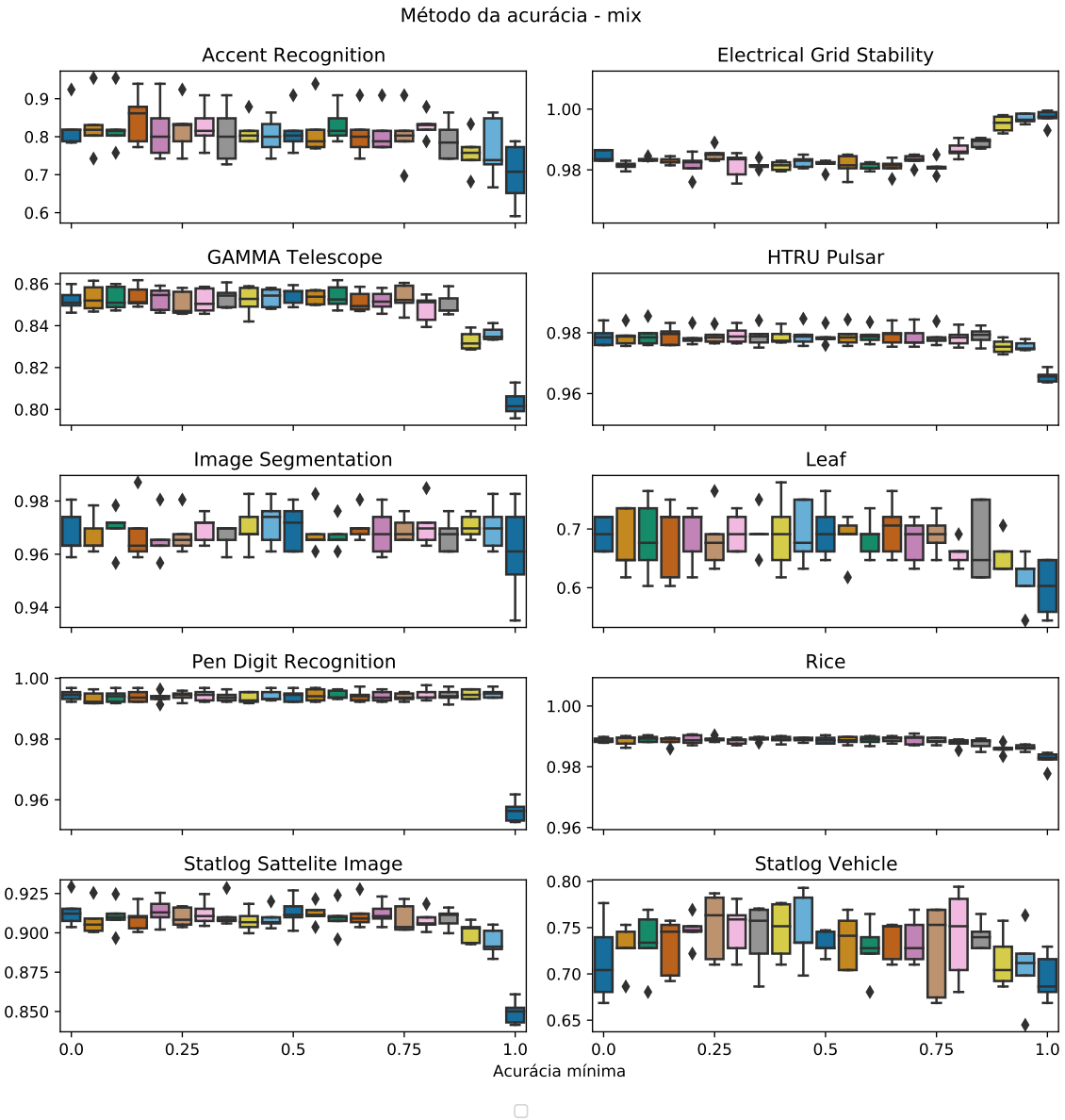

Figura B.2: Resultados do método de acurácia com classificador-base misto

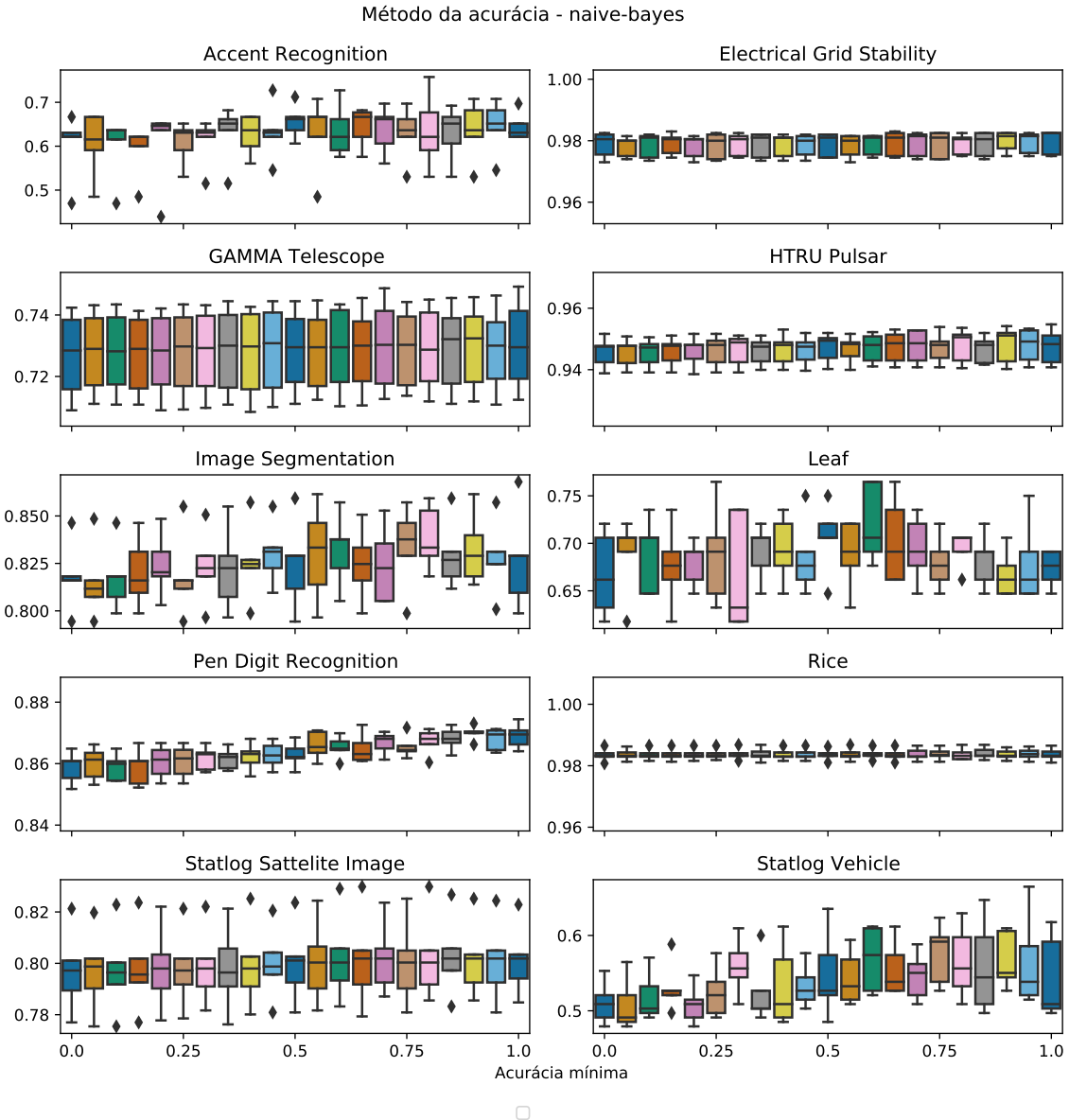

Figura B.3: Resultados do método de acurácia com classificador-base Naïve-Bayes

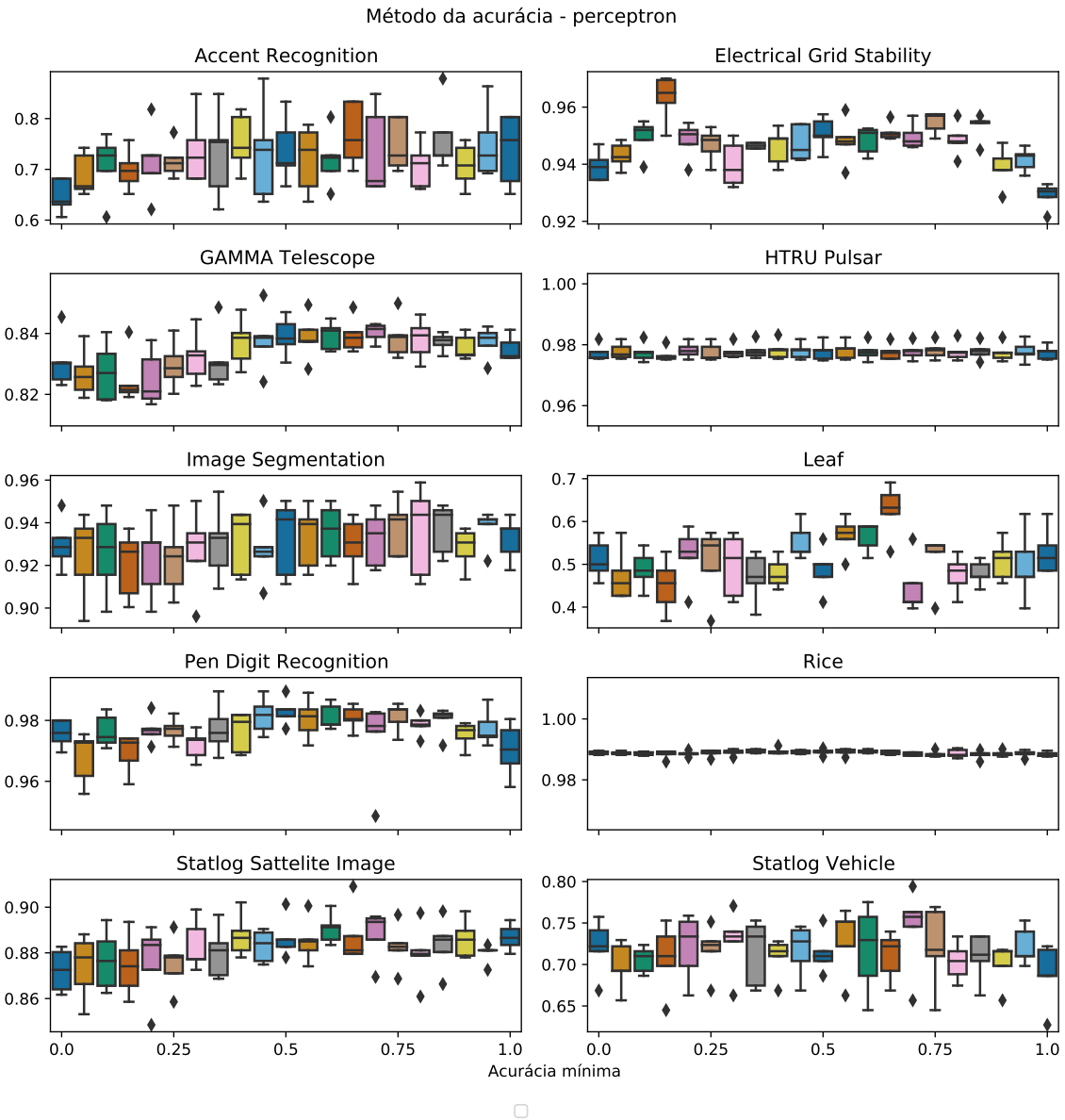

Figura B.4: Resultados do método de acurácia com classificador-base perceptron

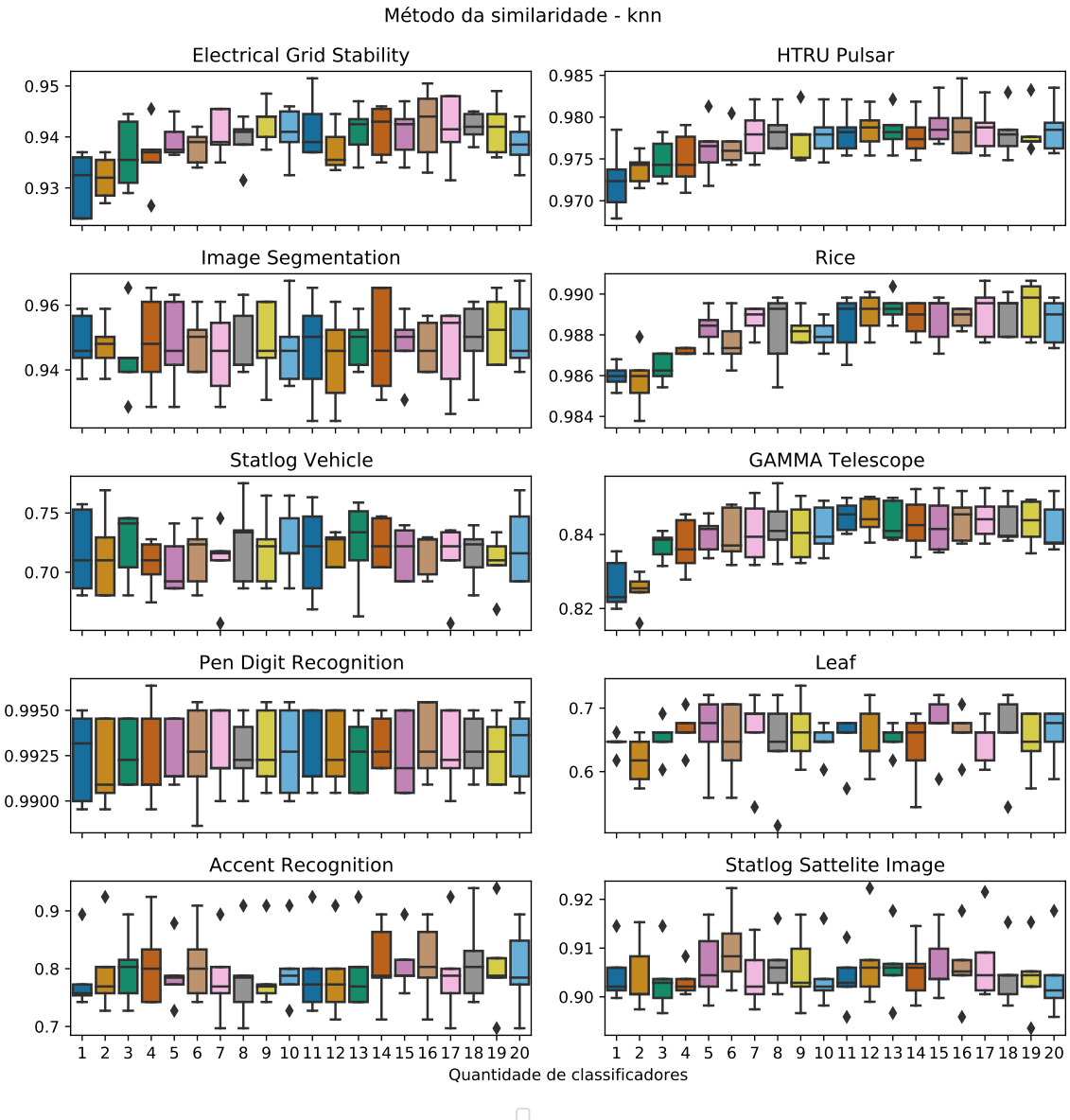

Figura B.5: Resultados do método de similaridade com classificador-base kNN

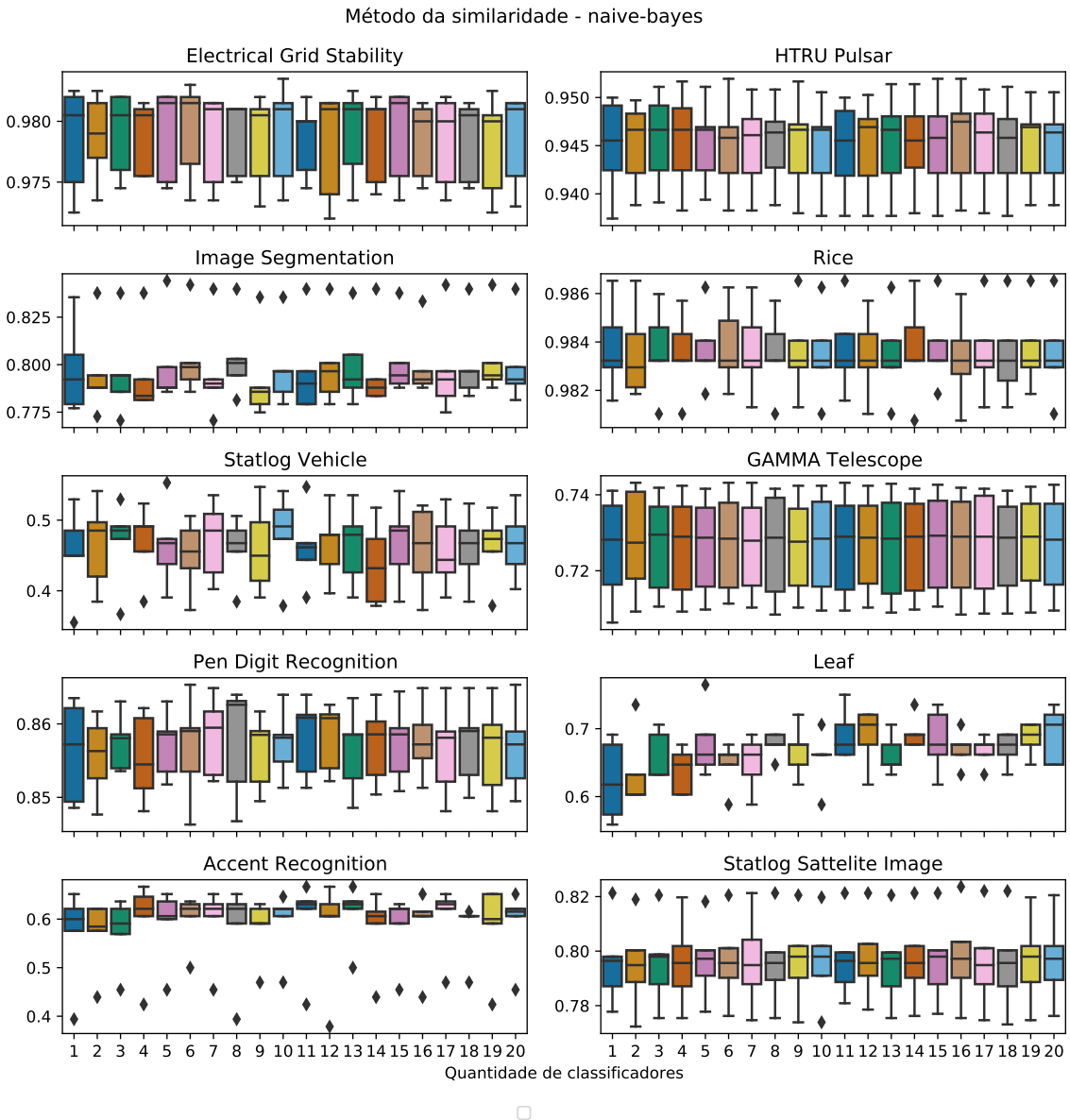

Figura B.6: Resultados do método de similaridade com classificador-base Naïve-Bayes

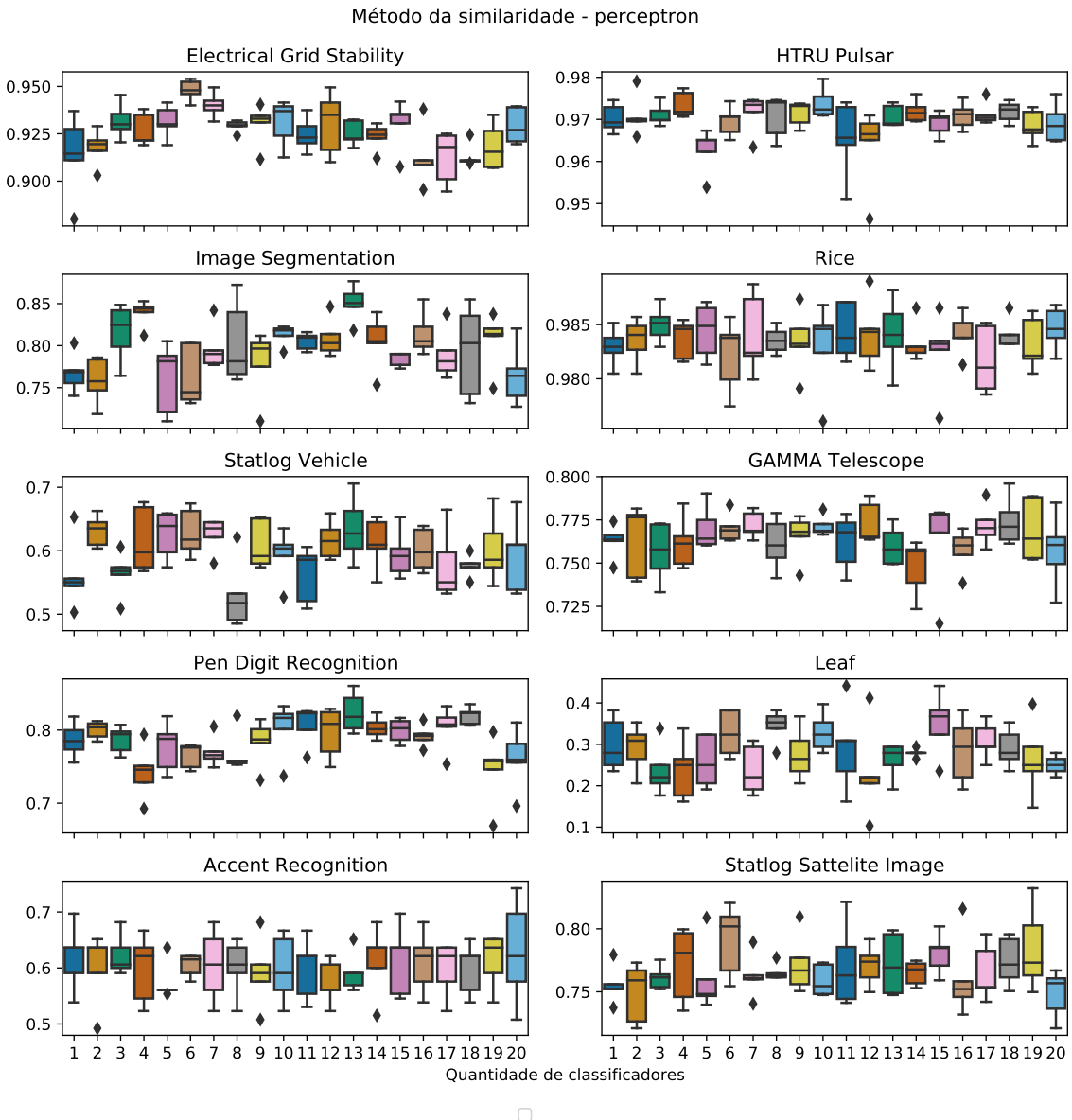

Figura B.7: Resultados do método de similaridade com classificador-base perceptron