Is P-value<0.05 Enough? Two Case Studies in Classifiers Evaluation

Nadine M. Neumann[1], Alexandre Plastino[1], Jony A. Pinto Junior[2], Alex A. Freitas[3]

[1]Instituto de Computação – Universidade Federal Fluminense
Niterói, RJ – Brasil

[2]Departamento de Estatística – Universidade Federal Fluminense
Niterói, RJ – Brasil

[3]School of Computing – University of Kent
Canterbury, Kent – England

{nadinemelloni,jarrais}@id.uff.br, plastino@ic.uff.br, A.A.Freitas@kent.ac.uk

Abstract. *A common tool used in the process of comparing classifiers is the statistical significance analysis, performed through the hypothesis test. However, there are many researchers attempting to obtain statistical significance through a blinding evaluating of the p-value<0.05 condition, ignoring important concepts such as the effect size and statistical power. This work highlights possible problems caused by the misuse of the hypothesis test and how the effect size and the statistical power can provide information for a better decision making. Therefore, two case studies applying Student's t-test and Wilcoxon signed-rank test for the comparison of two classifiers are presented.*

Resumo. *Uma ferramenta comumente utilizada no processo de comparação de classificadores é a análise da significância estatística, realizada através de teste de hipóteses. Entretanto, percebe-se que muitos pesquisadores estão buscando cegamente a significância estatística por meio da condição p-valor<0,05 e ignorando conceitos importantes como o tamanho do efeito e o poder do teste. Neste trabalho, são evidenciados possíveis problemas causados pelo mau uso dessa ferramenta e como o tamanho do efeito e o poder do teste acrescentam informação para uma melhor tomada de decisão. Para tanto, são apresentados dois estudos de caso com os testes t de Student e de Wilcoxon para a comparação de dois classificadores.*

1. Introdução

Nas áreas de Aprendizado de Máquina e Mineração de Dados, uma das tarefas mais importantes é a de Classificação, que permite determinar à qual classe determinado elemento pertence a partir dos valores dos seus atributos. A busca pelo melhor desempenho nessa tarefa faz com que novos algoritmos sejam propostos e, assim, é de fundamental importância que os pesquisadores tenham as ferramentas adequadas para comparar conscientemente as novas abordagens com as estratégias já existentes [Japkowicz and Shah 2011]. E uma ferramenta comumente utilizada nesse processo de comparação é a análise da significância estatística dos resultados.

A significância estatística é verificada por meio de algum teste de hipóteses que busca evidências para rejeitar uma hipótese conservadora, como por exemplo que dois classificadores têm resultados semelhantes. O resultado de um teste de hipóteses é, na maioria das vezes, dado por meio do *p-valor*, uma medida estatística útil, mas vem sendo utilizada abusivamente e mal interpretada [Wasserstein and Lazar 2016].

Outros conceitos de extrema importância, que têm sido ignorados pelos pesquisadores, são o poder do teste e o tamanho do efeito. O poder do teste é a probabilidade de se obter significância estatística para rejeitar a hipótese conservadora quando ela realmente é falsa e representa a adequação do teste ao contexto em que está sendo aplicado. Já o tamanho do efeito mede a força do resultado posto em teste, e ignorar essa medida é correr o risco de valorizar um resultado sem importância ou não considerar um resultado que poderia ser relevante.

Observa-se que a preocupação com essa questão está presente em diversas áreas. A Biologia [Nakagawa and Cuthill 2007], as Ciências do Esporte [Tomczak and Tomczak 2014], a Psicologia [Sharpe 2004] e a Medicina [Sullivan and Feinn 2012] são exemplos de onde já é valorizado o cálculo de uma medida de tamanho do efeito para acompanhar a análise de significância estatística.

Neste trabalho, foi feito um levantamento, no contexto de Aprendizado de Máquina e Mineração de Dados, considerando 11 artigos publicados em 2017 no periódico *Machine Learning*, que tratam de métodos de classificação e utilizam um ou mais testes de hipóteses para analisar seus resultados. Verificou-se que os pesquisadores da área estão simplificando a análise, resumindo-a a busca pelo $p\text{-}valor<0{,}05$ e nenhum artigo apresentou alguma medida do tamanho do efeito nem relatou o poder do teste realizado.

Para evidenciar a importância de uma boa análise, foram realizados dois estudos de caso para comparar o desempenho do classificador *k-nearest neighbors* (k-NN) com k=1 e k=3, ou simplesmente 1-NN e 3-NN, em bases de dados disponíveis no repositório da UCI [Dheeru and Karra Taniskidou 2017]. O k-NN, citado em [Wu et al. 2008] como um dos 10 algoritmos de Mineração de Dados mais influentes na comunidade acadêmica, tem como ideia principal determinar a classe do elemento de entrada como sendo a majoritária entre as classes dos seus k elementos mais semelhantes (k vizinhos mais próximos) na base de treinamento.

Nos estudos de caso, o desempenho dos classificadores foi medido por meio das acurácias obtidas, e a significância estatística da diferença entre as médias das acurácias foi verificada por meio dos testes t de Student e do teste de Wilcoxon, juntamente com o cálculo do tamanho do efeito e do poder do teste. Estes estudos ilustram como o *p-valor* e o tamanho do efeito podem levar a conclusões distintas no que diz respeito aos classificadores terem ou não resultados diferentes.

Este artigo está organizado da seguinte forma. Nas Seções 2 e 3, são apresentados os dois estudos de caso, um para o teste t de Student e outro para o teste de Wilcoxon, respectivamente, onde o *p-valor* e o tamanho do efeito discordam. Nessas seções, é discutido como seriam as conclusões tomadas com base em três análises: apenas no *p-valor*, com o *p-valor* e o tamanho do efeito e com as três medidas combinadas – *p-valor*, tamanho do efeito e poder do teste. Para finalizar, na Seção 4, são

apresentadas as conclusões deste trabalho, as práticas recomendadas para reduzir o risco de tomada de decisões equivocadas e possíveis direções para trabalhos futuros.

2. Primeiro Estudo de Caso: Teste t de Student

Nesta seção, será explorado um caso em que o teste de hipóteses não indica significância estatística, porém com um tamanho do efeito médio. Para tanto, será abordado um exemplo de aplicação do Teste t de Student para observações pareadas. Deseja-se verificar se os resultados dos classificadores 1-NN e 3-NN são diferentes utilizando-se as amostras de resultados observados a partir da base de dados *Mammographic Mass*, obtida no repositório da UCI.

Para a realização desse experimento, foi utilizada a Ferramenta Weka [Eibe Frank, Mark A. Hall, and Ian H. Witten 2016], onde estão implementados os classificadores adotados. Foi utilizado o método de validação cruzada com 10 partições. Em cada partição, foram aplicados o 1-NN e o 3-NN, ou seja, foram obtidas ao todo 10 pares de acurácias. Por esse motivo, deve-se utilizar testes de hipóteses para dados pareados. As acurácias obtidas estão apresentadas na Tabela 1.

Tabela 1. Acurácias obtidas por partição para cada classificador

	1	2	3	4	5	6	7	8	9	10
1-NN	77,32	71,88	72,92	73,96	71,88	70,83	78,12	72,92	81,25	81,25
3-NN	77,32	75,00	75,00	78,12	77,08	78,12	78,12	75,00	80,21	79,17

2.1. Conclusão com base no *p-valor*

Sejam X a variável de acurácias populacionais do 1-NN e Y a variável de acurácias populacionais do 3-NN. Para simplificar as hipóteses do teste, a variável $D = X - Y$ é definidia como a diferença populacional entre a acurácia do classificador A e acurácia do classificador B. Considere então (x_1,y_1), (x_2,y_2), $\ldots$, (x_n,y_n), onde $n = 10$, os dez pares de acurácias observadas e $d_i = x_i - y_i$ as diferenças entre esses pares de acurácias, onde $1 \leq i \leq n$. Inicialmente, é preciso ser verificada a suposição de normalidade das distribuições amostrais de X e Y para que o Teste t para dados pareados possa ser aplicado. Para isso, utiliza-se o teste de Kolmogorov-Smirnov para cada amostra, cuja hipótese nula é que as acurácias têm distribuição normal.

Sabe-se que as acurácias têm valores limitados entre 0 e 1, logo não têm distribuição normal. Mesmo com essa certeza, o teste de normalidade será realizado, pois a distribuição pode ser próxima da normal (simétrica com mesma média e mediana) e, com isso, não apresentar grandes perdas ao aplicar o teste t. O teste de Kolmogorov-Smirnov confirma a hipótese de normalidade da distribuição das acurácias do 1-NN e do 3-NN ao nível de significância de 5%, já que os p-valores obtidos são 0,68 e 0,83, respectivamente. Sendo assim, a hipótese nula (que representa a normalidade dos dados) não é rejeitada para ambos classificadores. Além dos testes de normalidade, existem outras maneiras de verificar se os dados seguem uma distribuição normal. Essa análise poderia ser feita, por exemplo, através da análise do gráfico Quantil-quantil, onde os quantis observados são comparados aos quantis teóricos da distribuição normal.

Para definir as hipóteses a serem testadas, é necessário decidir se será um teste unilateral ou bilateral. Como não há o desejo de verificar se um classificador tem resultado melhor que o outro, porém apenas se eles têm resultados diferentes, será aplicado o teste bilateral. Dessa forma, a hipótese nula é de que a média populacional das acurácias obtidas pelo classificador 1-NN é igual à média populacional das acurácias obtidas pelo classificador 3-NN na base em análise. E a hipótese alternativa é de que a média das acurácias populacionais obtidas por esses classificadores são diferentes. Sendo assim, a hipótese nula pode considerar que que não existe diferença entre as médias das acurácias populacionais obtidas pelos classificadores 1-NN e 3-NN na base em análise, e a hipótese alternativa que essa diferença populacional é diferente de zero. Ou seja:

$$\begin{cases} H_0: & \mu_D = 0 \\ H_1: & \mu_D \neq 0 \end{cases} \tag{1}$$

onde μ_D é a diferença entre as médias das acurácias populacionais dos classificadores.

A diferença entre as médias observadas das acurácias amostrais é $\overline{d} = -2{,}08$. Deseja-se verificar se a diferença populacional é estatisticamente significante (se é diferente de 0), assim como foi observada na amostra, ou se essa diferença observada é particularidade daquelas amostras e não é uma característica da população.

A estatística do teste é calculada através da expressão

$$T = \frac{\overline{D} - \mu_D}{\sqrt{\frac{S_D^2}{n}}} \tag{2}$$

onde n é o tamanho da amostra e S_D^2 é a variância das diferenças populacionais, que é estimado pela variância das diferenças amostrais. Com H_0 verdadeira, ou seja, $\mu_D = 0$, T segue uma distribuição t de Student com $n-1$ graus de liberdade. Essa distribuição só é conhecida se D tem distribuição normal, por isso é necessário a verificação inicial da normalidade.

Então,

$$t = \frac{\overline{d} - 0}{\frac{s_d}{\sqrt{n}}} = \frac{-2{,}08}{\frac{2{,}95}{\sqrt{10}}} = -2{,}24. \tag{3}$$

O *p-valor* é a probabilidade de se obter uma estatística de teste igual ou mais extrema quanto a observada na amostra, assumindo verdadeira a hipótese nula [Fisher 1925]. Sendo assim, no teste bilateral, o *p-valor* é calculado como

$$p\text{-}valor = P(T \geq |t|) + P(T \leq -|t|). \tag{4}$$

Então, calculando o *p-valor*, obtém-se $p\text{-}valor = 0{,}0261 + 0{,}0261 = 0{,}0522$. Como o *p-valor* é maior que o nível de significância $\alpha = 0{,}05$, o teste t para amostras pareadas não obteve evidências para rejeitar a hipótese nula que é a hipótese conservadora de que a média das acurácias amostrais obtidas pelos classificadores 1-NN e 3-NN são iguais. Sendo assim, ao nível de significância de 5% não é possível afirmar que os classificadores 1-NN e 3-NN têm médias de acurácias diferentes.

2.2. Avaliando o Tamanho do Efeito

Uma medida de tamanho do efeito pode complementar a conclusão tirada com base no *p-valor*. O tamanho do efeito pode ser definido como o grau em que o fenômeno está presente na população, isto é, a diferença efetiva na população [Cohen 1988]. Assim, quanto maior for o tamanho do efeito, maior será a manifestação do fenômeno (diferença) na população. Existem diversas medidas de tamanho do efeito, e segundo Joseph Edward [Conboy 2012] o uso destas métricas é uma alternativa ao conceito de significância estatística, tratando de noções de significância prática específica.

Além disso, outra motivação para o cálculo de uma medida do tamanho do efeito é que o nível de significância é afetado por diversas características do estudo, sendo o tamanho da amostra o mais determinante [Snyder and Lawson 1993]. Assim, é mais provável obter um *p-valor* significativo com tamanhos grandes de amostras e inversamente, em amostras pequenas, o *p-valor* pode não ser significativo [Santo and Daniel 2015].

O objetivo desse trabalho não inclui discutir sobre as diversas medidas de tamanho do efeito que existem, por isso, serão apresentadas apenas as medidas aqui utilizadas para os testes t e de Wilcoxon, ambos para amostras pareadas. Para o Teste t será utilizado o *d de Cohen* que é calculado da seguinte maneira [Cohen 1988]:

$$d'_{cohen} = \left| \frac{\overline{d}}{s_d} \right|. \tag{5}$$

Pela Equação 3, é possível reescrever o *d de Cohen* como:

$$d'_{cohen} = \left| \frac{\frac{\overline{d}}{\sqrt{n}}}{\frac{s_d}{\sqrt{n}}} \right| = \left| \frac{t}{\sqrt{n}} \right|. \tag{6}$$

Logo, para o caso em questão, a medida d de Cohen, é calculada da seguinte forma $d'cohen = \frac{|t|}{\sqrt{n}} = \frac{2{,}24}{\sqrt{10}} = 0{,}71$, onde t é a estatística de teste e n é a quantidade de pares em comparação (tamanho da amostra).

Seguindo a sugestão de Cohen, o tamanho do efeito calculado $d'cohen = 0{,}71$ representa um tamanho do efeito médio. Ou seja, o teste de hipóteses não indica diferença estatisticamente significativa entre as médias das acurácias dos classificadores 1-NN e 3-NN, porém a diferença entre essas médias tem tamanho do efeito médio indicando que a magnitude dessa diferença pode ser importante. Sendo assim, o *p-valor* e o tamanho do efeito indicam resultados diferentes. Nesse caso, o cálculo do poder do teste, como será visto na subseção seguinte, acrescenta informações necessárias para uma tomada de decisão mais fundamentada.

2.3. Avaliando o Poder do Teste

O poder do teste representa a probabilidade do teste rejeitar H_0 quando H_0 realmente é falsa, ou seja, é a probabilidade do teste afirmar que os resultados dos classificadores 1-NN e 3-NN são diferentes para a base de dados quando realmente são. Na realidade, o poder do teste é uma função pois, sendo $H_0 : \mu_D = 0$ falsa, não

se sabe o valor verdadeiro para μ_D, sabe-se apenas que ele é diferente de zero, logo o poder do teste é calculado para todos os valores possíveis de μ_D.

Considerando que o valor real da diferença entre os classificadores é igual à diferença observada ($-2{,}08$), o poder do teste t é 0,49. Ou seja, a probabilidade de o teste t afirmar que os classificadores são diferentes se a diferença real for $-2{,}08$ é de apenas 49%. Se o tamanho da amostra fosse aumentado, por exemplo, para $n = 25$ e $n = 50$, o poder do teste seria 92% e 98%, respectivamente.

Conforme visto anteriormente, o teste t não obteve evidências para rejeitar a hipótese nula, já que *p-valor*=0,052, e assim não foi possível afirmar que as médias das acurácias populacionais dos classificadores 1-NN e 3-NN são diferentes ao nível de significância de 5%. Porém, o tamanho do efeito médio ($d'cohen = 0{,}71$) para a diferença entre as acurácias indica que a magnitude dessa diferença pode ser um resultado importante. Ou seja, as duas medidas podem levar a conclusões distintas no que diz respeito aos classificadores terem ou não resultados diferentes.

O cálculo do poder do teste permite compreender o possível motivo das medidas anteriores terem resultados distintos. O poder do teste de apenas 49%, indica que o teste tem baixa probabilidade de afirmar que os classificadores são diferentes quando essa diferença é igual a $-2{,}08$. Ou seja, o teste t foi aplicado mesmo sendo pouco poderoso para essa amostra. Portanto, esse exemplo ilustra não somente a importância do cálculo das três medidas, mas também o risco de tomar uma decisão equivocada se for utilizado apenas o *p-valor*. Além disso, ilustra o equívoco que é aplicar um teste de hipóteses sem conhecer o poder desse teste para a amostra.

3. Segundo Estudo de Caso: Teste de Wilcoxon

Nesta Seção, será explorado um caso em que o teste de hipóteses indica significância estatística, porém com um tamanho do efeito pequeno. Para tanto, será abordado um exemplo de aplicação do Teste de Wilcoxon para observações pareadas. Deseja-se verificar se os resultados dos classificadores 1-NN e 3-NN são diferentes para a base *Wholesale3*, também obtida no repositório da UCI. Outra diferença em relação ao exemplo anterior é que, no lugar do método de validação cruzada com 10 partições, foram utilizadas 30 partições, ou seja, a amostra nesse exemplo tem tamanho 30. Com uma amostra maior, aumenta-se a chance de o teste indicar significância estatística e de se obter o caso a ser ilustrado.

As acurácias obtidas pelos classificadores 1-NN e 3-NN em cada uma das 30 partições são apresentadas na Tabela 2.

3.1. Conclusão com base no *p-valor*

Sejam X a variável de acurácias populacionais do 1-NN e Y a variável de acurácias populacionais do 3-NN. Para simplificar as hipóteses do teste, é possível definir $D = X - Y$, ou seja, a variável D é a diferença populacional entre a acurácia do classificador A e acurácia do classificador B. Considere então (x_1,y_1), (x_2,y_2), ..., (x_n,y_n), onde $n = 30$, os 30 pares de acurácias observadas e as diferenças $d_i = x_i - y_i$, onde $1 \leq i \leq n$. Será aplicado o teste bilateral de Wilcoxon para amostras pareadas considerando nível de significância de 5%, cujas hipóteses são apresentadas em (7),

Tabela 2. Acurácias de cada partição obtidas pelos classificadores 1-NN e 3-NN

	1-NN	3-NN		1-NN	3-NN
1	80,00	86,67	16	86,67	86,67
2	93,33	93,33	17	73,33	80,00
3	93,33	86,67	18	93,33	93,33
4	93,33	93,33	19	73,33	86,67
5	86,67	93,33	20	86,67	86,67
6	93,33	100,00	21	92,86	92,86
7	86,67	93,33	22	78,57	92,86
8	93,33	93,33	23	85,71	78,57
9	80,00	86,67	24	71,43	78,57
10	86,67	93,33	25	100,00	85,71
11	93,33	93,33	26	85,71	85,71
12	100,00	100,00	27	92,86	92,86
13	80,00	100,00	28	85,71	85,71
14	86,67	93,33	29	78,57	92,86
15	86,67	93,33	30	85,71	85,71

onde δ_D é a mediana da diferença entre as acurácias populacionais dos classificadores 1-NN e 3-NN na base em análise.

$$\begin{cases} H_0: & \delta_D = 0 \\ H_1: & \delta_D \neq 0 \end{cases} \tag{7}$$

O *p-valor* para o teste de Wilcoxon é 0,025 e, por ser menor que o nível de significância $\alpha = 0{,}05$, indica que há evidências para rejeitar a hipótese nula. Portanto, é possível afirmar, ao nível de significância de 5%, que as medianas das acurácias populacionais dos classificadores são estatisticamente diferentes, ou seja, as acurácias dos classificadores diferem em localização.

3.2. Avaliando o Tamanho do Efeito

Uma medida de tamanho do efeito pode complementar a conclusão tirada com base no *p-valor*. Para o caso em questão (teste de Wilcoxon para amostras pareadas) será utilizada a medida r proposta por Cohen [Cohen 1988]:

$$r = \frac{z}{\sqrt{2n}} \tag{8}$$

onde z é a estatística do teste de Wilcoxon com aproximação pela normal.

Logo, o tamanho do efeito é calculado da seguinte forma $r = \frac{|z|}{\sqrt{2n}} = \frac{2{,}24}{\sqrt{60}} = 0{,}28$. Seguindo a classificação proposta por Cohen, tem-se que $r = 0{,}28$ representa um tamanho do efeito pequeno. Ou seja, o teste de hipóteses indica que a diferença entre as medianas das acurácias populacionais dos classificadores 1-NN e 3-NN é estatisticamente significativa, porém o tamanho do efeito pequeno indica que a magnitude dessa diferença pode não ser um resultado importante ou relevante para o pesquisador. Sendo assim, o *p-valor* e o tamanho do efeito indicam resultados diferentes. Nesse caso, o cálculo do poder do teste, como será visto na

subseção seguinte, acrescenta informações necessárias para uma tomada de decisão mais fundamentada.

3.3. Avaliando o Poder do Teste

Para realizar o Teste de Wilcoxon não foi necessário fazer nenhum pressuposto sobre a distribuição da população, uma vez que se trata de um teste não paramétrico. Porém, para o cálculo do poder do teste, é necessário que a distribuição da diferença entre as acurácias populacionais seja conhecida, o que torna esse cálculo diferente do que foi para o Teste t, uma vez que ele precisa ser obtido através de simulação [Coelho Barros and Mazucheli 2005].

Foram simulados 1000 pares de amostras com distribuição normal e de tamanho 30, com os parâmetros média e desvio padrão respectivos das amostras X e Y. Para cada par de amostra, foi realizado o Teste de Wilcoxon e feito a proporção de quantos foram significativos entre o total. Foram obtidos os seguintes resultados: dos 1000 pares de amostras, 433 foram significativos para o Teste de Wilcoxon, portanto, pode-se dizer que o poder do teste é de aproximadamente 43% para uma diferença real igual à diferença amostral de $\overline{d} = -3{,}36$. Ou seja, a probabilidade do teste afirmar que os classificadores são distintos, quando essa diferença for $-3{,}36$, é de aproximadamente 43%.

Conforme visto anteriormente, é possível afirmar que as acurácias populacionais dos classificadores 1-NN e 3-NN são diferentes em localização, ou seja, têm medianas diferentes, ao nível de significância de 5%, já que *p-valor*=0,025 no teste de Wilcoxon. Porém, o tamanho do efeito pequeno ($r = 0{,}28$) indica que a magnitude dessa diferença é pequena e possivelmente pode ser um resultado sem relevância. Ou seja, as duas medidas podem levar a conclusões distintas no que diz respeito aos classificadores terem ou não resultados diferentes.

O cálculo do poder do teste permite compreender o possível motivo das medidas anteriores terem resultados distintos. O poder do teste de apenas 43%, indica que o teste tem baixa probabilidade de afirmar que os classificadores são diferentes quando essa diferença populacional for igual a diferença amostral. Ou seja, o teste de Wilcoxon foi aplicado mesmo sendo pouco poderoso para essa amostra. Portanto, assim como no exemplo do teste t, esse exemplo ilustra não somente a importância do cálculo das três medidas, mas também o risco de tomar uma decisão equivocada se for utilizado apenas o *p-valor*. Além disso, ilustra o equívoco que é aplicar um teste de hipóteses sem conhecer o poder desse teste para a amostra em análise.

4. Conclusões e Práticas Recomendadas

Neste trabalho, foram realizados os dois estudos de caso nos quais foi mostrado como a tomada de decisão com base apenas no *p-valor* é um procedimento irresponsável, já que pode levar a valorizar um resultado sem relevância ou a ignorar um resultado importante. Além disso, foi evidenciado como o cálculo do tamanho do efeito e do poder do teste auxilia na comparação de dois classificadores aplicados a uma base de dados.

O primeiro estudo de caso avaliou se os classificadores 1-NN e 3-NN têm resultados diferentes, em determinada base de dados, utilizando o teste t com uma

amostra de 10 pares de acurácias. Não foi obtida significância estatística entre as diferenças das acurácias, porém essa diferença tinha tamanho do efeito médio. Com cálculo do poder do teste, foi verificado que o teste aplicado era pouco poderoso, ou seja, o pesquisador poderia desistir de um resultado com magnitude importante se considerasse apenas o *p-valor*, e sua conclusão seria baseada em um teste fraco.

Em casos como este, quando o teste realizado é pouco poderoso, o pesquisador tem a oportunidade de tentar aumentar o poder do teste buscando mais elementos da amostra. Ou seja, um pesquisador não deve desistir do seu estudo por não ter encontrado significância estatística por meio de um teste com poder baixo, pois pode representar a perda de um resultado importante, como indicado pelo tamanho do efeito.

No segundo estudo de caso, o objetivo era o mesmo: comparar os classificadores 1-NN e 3-NN em uma determinada base de dados. Porém, foi utilizado o teste de Wilcoxon sobre uma amostra de 30 pares de acurácias. O teste obteve significância estatística entre a diferença da mediana das acurácias, porém tanto o tamanho do efeito quanto o poder do teste foram baixos. Ou seja, o *p-valor* e o tamanho do efeito indicaram resultados diferentes e o teste de Wilcoxon foi aplicado mesmo sendo pouco poderoso para essa amostra. Portanto, esse exemplo novamente ilustrou não somente a importância do cálculo das três medidas, mas também o risco de tomar uma decisão equivocada se for utilizado apenas o *p-valor*.

A análise de significância estatística deve ser realizada de maneira muito consciente pelo pesquisador, já que, um teste de hipóteses mal aplicado pode levar a graves erros de conclusão. O cálculo do poder do teste e de alguma medida do tamanho do efeito colaboram para que a tomada de decisão seja mais fundamentada e responsável.

Portanto, a aplicação de um teste de hipóteses sempre deve ser acompanhada do cálculo do poder do teste. Além disso, o tamanho do efeito pode complementar a conclusão tirada com base no *p-valor*. Entretanto, como visto no levantamento realizado neste estudo e descrito na Seção 1, com base nos artigos publicados em 2017 no periódico *Machine Learning*, nenhum dos 11 artigos que foram verificados apresentou o poder do teste ou alguma medida de tamanho do efeito juntamente com o teste de hipóteses realizado.

Com o objetivo de ampliar o estudo realizado neste trabalho, pretende-se, como trabalhos futuros, analisar um conjunto maior de resultados (empíricos e simulados) obtidos a partir de diversos tamanhos de amostras, diferentes bases de dados, utilizando um conjunto maior de classificadores. Pretende-se também analisar outros testes de hipóteses paramétricos e não paramétricos, como por exemplo o teste de Friedman e ANOVA, que teriam como objetivo, no contexto deste estudo, comparar o desempenho de diferentes classificadores para um conjunto de bases de dados. Com isso, será possível verificar, no processo de comparação de classificadores, a frequência dos casos que podem levar a decisões equivocadas quando as análises são feitas com base apenas no *p-valor*. Dessa forma, espera-se estimular a apresentação do poder do teste e de alguma medida de tamanho do efeito nas pesquisas das áreas de Aprendizado de Máquina e Mineração de Dados, a fim de

tornar as conclusões estatísticas mais fundamentadas.

Referências

Coelho Barros, E. A. and Mazucheli, J. (2005). Um estudo sobre o tamanho e poder dos testes t-Student e Wilcoxon. *Acta Scientiarum: Technology*, 27(1):23–32.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: erlbaum, 2nd edition.

Conboy, J. E. (2012). Algumas medidas típicas univariadas da magnitude do efeito. *Análise Psicológica*, 21(2):145–158.

Dheeru, D. and Karra Taniskidou, E. (2017). UCI machine learning repository.

Eibe Frank, Mark A. Hall, and Ian H. Witten (2016). *The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques"*. Morgan Kaufmann, 4th edition.

Fisher, R. A. (1925). *Statistical methods for research workers*. Springer.

Japkowicz, N. and Shah, M. (2011). *Evaluating learning algorithms: a classification perspective*. Cambridge University Press.

Nakagawa, S. and Cuthill, I. C. (2007). Effect size, confidence interval and statistical significance: a practical guide for biologists. *Biological Reviews*, 82(4):591–605.

Santo, H. E. and Daniel, F. B. (2015). Calcular e apresentar tamanhos do efeito em trabalhos científicos (1): As limitações do p<0,05 na análise de diferenças de médias de dois grupos. *Revista Portuguesa de Investigação Comportamental e Social*, 1(1):3–16.

Sharpe, D. (2004). Beyond significance testing: Reforming data analysis methods in behavioral research. 45(4):317–319.

Snyder, P. and Lawson, S. (1993). Evaluating results using corrected and uncorrected effect size estimates. *The Journal of Experimental Education*, 61(4):334–349.

Sullivan, G. M. and Feinn, R. (2012). Using effect size-or why the p-value is not enough. *Journal of Graduate Medical Education*, 4(3):279–282.

Tomczak, M. and Tomczak, E. (2014). The need to report effect size estimates revisited. an overview of some recommended measures of effect size. *Trends in Sport Sciences*, 21(1):19–25.

Wasserstein, R. L. and Lazar, N. A. (2016). The ASA's statement on p-values: context, process, and purpose. *The American Statistician*.

Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Ng, A., Liu, B., Philip, S. Y., et al. (2008). Top 10 algorithms in data mining. *Knowledge and information systems*, 14(1):1–37.