Estudo de Caso 02: Comparação entre Algoritmos de Classificação

Equipe 04

15 de maio de 2017

Coordenador: Danny Tonidandel
Relator: Alessandro Cardoso
Verificador: Gustavo Vieira
Monitor: Bernardo Marques

O Experimento

Este estudo de caso consiste na comparação de algoritmos de classificação em que um deles consiste no *padrão atual* e o outro em um novo algoritmo proposto, que utilizada uma técnica de simplificação baseada em inferência estatística.

Segundo os pesquisadores responsáveis pelo algoritmo proposto, este apresenta uma melhora significativa frente ao padrão atual com relação ao tempo requerido para a classificação e que o essa nova abordagem não resulta em grandes perdas de desempenho em termos de acurácia da classificação. Assim, busca-se responder:

1. O método proposto realmente apresenta ganhos em relação ao tempo de execução, quando comparado ao método padrão?
2. O método proposto não resulta em variações consideráveis de acurácia?

Para que sejam investigados os questionamentos acima são desejadas as seguintes características para os testes estatísticos:

Nível de significância: $\alpha = 0.05$;

Tamanho de efeito de interesse prático para os ganhos de tempo: $d^{\star}_{tempo} = 1.0$;

Margem de não-inferioridade para acurácia: $\delta^{*}_{acuracia} = 0.05$.

Potência desejada: $\pi = 0.8$;

Planejamento experimental

Os dados experimentais para os testes foram gerados por meio de uma aplicação disponibilizada na web em *http* : *//orcslab.cpdee.ufmg.br/*3838*/classdata*. Os dados são referentes ao tempo necessário para classificação (*Time.s*) e acurácia (*Accuracy*) por cada um dos algoritmos em cada instância e em cada execução. Para geração dos dados o grupo de trabalho deve fornecer a data de nascimento do membro mais jovem da equipe e selecionar um número de instâncias e execuções por instâncias.

Estes dados são apresentados conforme seleção de número de instâncias e número de repetições das mesmas. Segundo [1] quando cada par de observação é coletado sobre condições homogêneas, ainda assim estas podem variar de um par para outro. Conforme apresentado em [2], a variabilidade devida aos diferentes problemas de teste é uma forte fonte de variação espúria que pode e deve ser controlada e que uma solução elegante para eliminar a influência deste inconveniente é o pareamento das medições por instância. O número de execuções por instância escolhido pelo grupo foi 30, por se tratar de um número comumente utilizado.

Conforme a apresentação das amostras, verificou-se a necessidade de efetuar os testes estatísticos com as amostras pareadas. Para [1] o procedimento experimental mais adequado quando os dados são coletados aos pares é o *t-test pareado*. Para tal, seja (X_{11}, X_{21}), (X_{12}, X_{22}),..., (X_{1n}, X_{2n}) um conjunto de n observações pareadas onde assumimos que a média e a variância da população representada por X_1 são μ_1 e σ_1^2, e a média e a variância da população representada por X_2 são μ_2 e σ_2^2. Defini-se as diferenças entre cada par de observações como $D_j = X_{1j} - X_{2j}$, j= 1, 2, p,... n. Os D_j's são assumidos como sendo normalmente distribuídos na média.

Definicão de hipóteses

Os testes de hipótese elaborados são:

- Referente ao questionamento 1, definiu-se o teste de hipótese unilateral convencional, em que a hipótese nula é de que o algoritmo proposto apresenta tempo de execução similar ao algoritmo atual. Já a hipótese alternativa é a de que o algoritmo proposto é mais rápido que o padrão atual. Esta formulação é apresentada abaixo.

$$\begin{cases} H_0 : \mu_{pt} - \mu_{at} = 0 \\ H_1 : \mu_{pt} - \mu_{at} < 0 \end{cases}$$

Os valores de μ_{pt} e μ_{at} são as médias de tempo de execução dos algoritmos *proposto* e *padrão atual*, respectivamente.

- Referente ao questionamento 2, definiu-se o teste de não-inferioridade, em que a hipótese nula afirma que o algoritmo proposto apresenta acurácia média inferior ao algoritmo atual. A hipótese alternativa afirma que o algoritmo proposto não apresenta acurácia média inferior ao padrão atual:

$$\begin{cases} H_0 : \mu_{pa} - \mu_{aa} = -\delta^*_{acurácia} \\ H_1 : \mu_{pa} - \mu_{aa} > -\delta^*_{acurácia} \end{cases}$$

Os valores de μ_{pa} e μ_{aa} são os valores arbitrários de acurácia dos algoritmos *proposto* e *padrão atual* respectivamente e $\delta^*_{acurácia}$ é a margem de não-inferioridade para a acurácia.

Tamanho amostral

No caso de amostras pareadas, entende-se por amostras, o números de instâncias conforme exposto por [3], ou problemas a serem resolvidos pelos algoritmos. Outra consideração é o número de repetições a serem adotadas por instância. Para este estudo adotou-se empiricamente o valor de 30 repetições por instância, seguindo recomendação expressas em [2].

Para a determinação do tamanho amostral foi realizado um *Estudo Piloto*, de forma a determinar uma estimativa inicial para os testes atendendo aos parâmetros desejados, obtendo uma cardinalidade de 800, para um erro relativo de $e_n = 0.10$, em $n_{pilot} \approx 2\left(\frac{z_{\alpha_n/2}}{e_n}\right)^2$. Por fornecer um número muito maior do que o número de amostras necessárias para um *Estudo Piloto*, a equipe decidiu por adotar um método iterativo. Como $d^\star_{tempo}$ é conhecido, ele foi utilizado para o cálculo, segundo a equação:

$$n = 2\left(\frac{1}{d^\star_{tempo}}\right)^2 (t_{\alpha/2} + t_\beta)^2 .$$

Os valores de $t_{\alpha/2}$ e t_β foram subtituidos para $z_{\alpha/2}$ e z_β na equação anterior, considerando-se a prmeira iteração, de forma a fazê-los independetes de n. As iterações seguintes convergiram rapidamente para um valor de 17 amostras, utilizadas no *Estudo Piloto* conforme as entradas:

Nascimento: 21/11/1992;

Número de instâncias de cada problema: 17;

Número de execuções: 30;

Por meio dos dados do *Estudo Piloto* obteve-se o desvio padrão das amostras dos algoritmos, que foram, por sua vez, utilizados na obtenção do tamanho amostral mais adequado aos testes finais, Utilizando a função *calcN_tost2* fornecida por [2].

Os testes definitivos serão feitos utilizando-se 33 instâncias de problemas, que permitirá, por sua vez, a geração de um novo conjunto de dados *1992-11-21_33_30.csv*, em que:

Para a geração e coleta dos dados foi adotado o maior tamanho amostral verificado para o tempo de execução e acurácia. Neste caso, quem apresentou o maior valor foi o tempo de execução. Este valor atende às características exigidas para ambos os testes sem necessidade de uma nova coleta de dados para o caso da acurácia. Os dados foram gerados conforme abaixo:

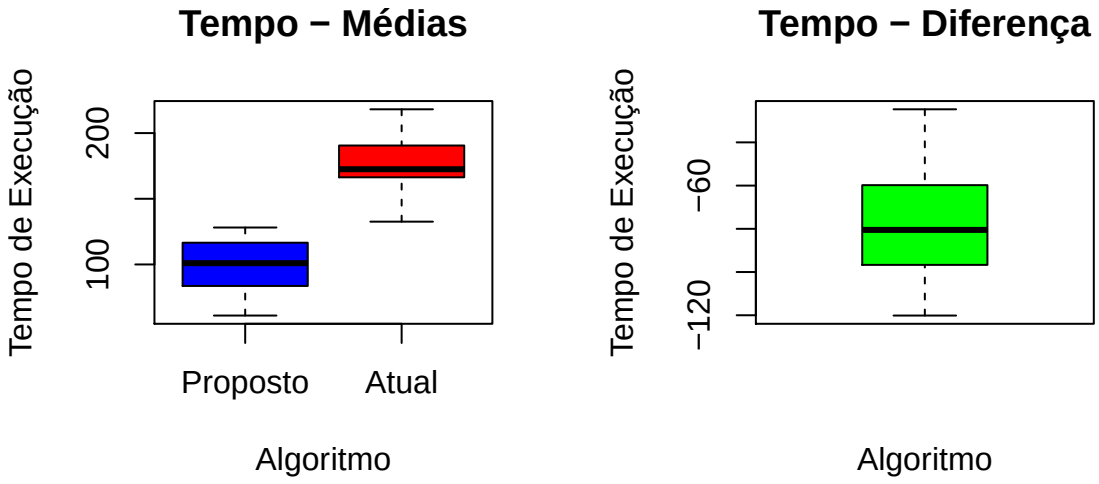
- Nascimento: 21/11/1992;
- Número de instâncias de cada problema: 33;
- Número de execuções: 30;

Teste de Hipóteses

Problema 1

Com o resultado do teste de hipóteses observa-se que não há evidências suficientes para se aceitar a hipótese nula (teste unilateral emparelhado), considerando-se a estatística gerada e o valor p para o nível de significância e graus de liberdade definidos. Assim, conclui-se que o algoritmo proposto apresenta um desempenho melhor, em relação às médias dos tempos de execução, se comparado ao algoritmo padrão.

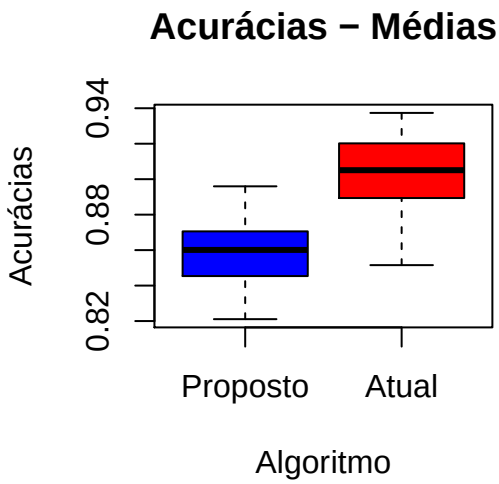
O gráfico bloxplot abaixo evidencia a diferença entre as médias dos tempos de execução. Podemos ver que a média é menor que zero, sugerindo que o algoritmo proposto possui média de tempo de execução menor que a média do algoritmo padrão.

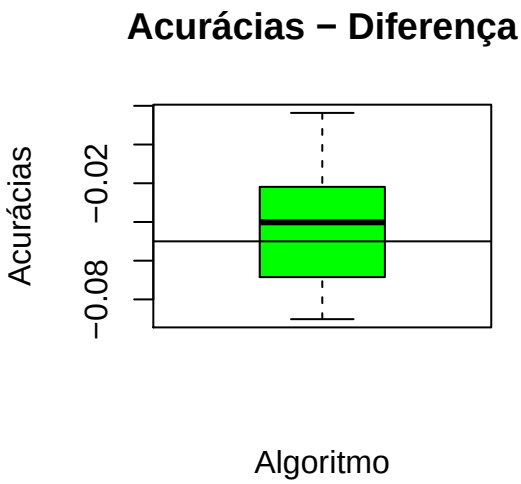


##

Problema 2 O teste de hipóteses referente ao questionamento 2 foi implementado e seu resultado indica que que não existem evidências suficientes que levem à rejeição da hipótese nula, para os requisitos estabelecidos.

O gráfico bloxplot referente ao experimento 2 evidencia a diferença entre as médias das acurácias. Podemos ver que o intervalo de significância não está completamente acima da margem de não-inferioridade da acurácia dada. Por tanto não foi possível estabelecer a não-inferioridade do algoritmo em relação à sua acurácia.



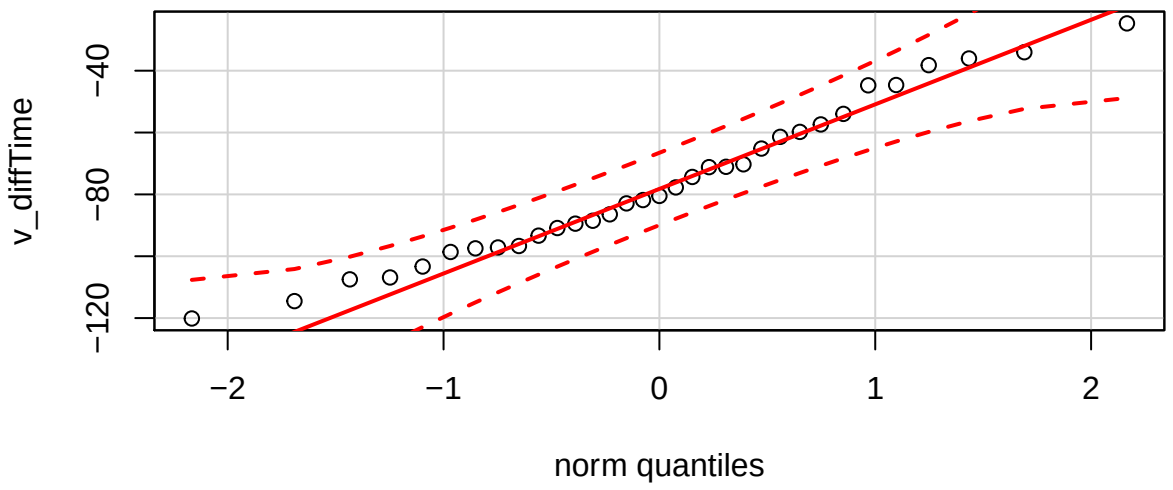


Validação

Problema 1

A premissa de normalidade para o teste do questionamento 1 foi testada inicialmente de forma visual a partir do gráfico QQplot e pelo teste de Shapiro-Wilk. Ambos sugerem a normalidade dos dados.

```
## Hypothesis validation - Time
qqPlot(v_diffTime)
```

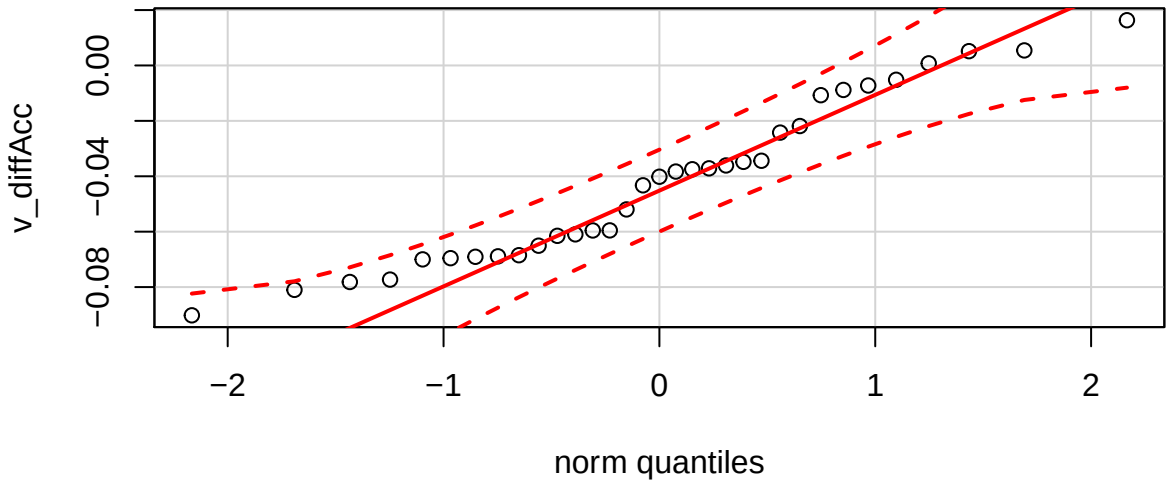



```
v_shapiroTime = shapiro.test(v_diffTime)
v_dwTime = dwtest(v_diffTime~1)
```

Problema 2

A premissas de normalidade e independência também sugerem que a amostra provêm de uma distribuição normal, com autocorrelação zero (independentes):

```
## Hypothesis validation - Acuracy
qqPlot(v_diffAcc)
```



```
v_shapiroAcc = shapiro.test(v_diffAcc)
v_dwAcc = dwtest(v_diffAcc~1)
```

Conclusão

Pode-se afirmar que o método proposto tem ganhos significativos em termos dos tempos de execução. Em relação ao segundo resultado, como não foi possível rejeitar a hipótese nula com o nível de significância desejado, não é possível afirmar concluir que o método proposto é não-inferior ao método padrão.

Vale ainda ressaltar que o tamanho amostral necessário para as características desejadas do teste foram consideravelmente distoantes, a saber, 33 para o primeiro teste e 4 para o segundo. Foi então decidido fazer apenas um teste com o número mais alto de amostras, uma vez que isto não implicaria em perdas para nenhum dos casos, e não seria necessário fazer mais de uma "coleta". Um resultado disto foi que para o segundo teste a potência calculada para o teste foi 1. Isto significa que a chance de ocorrer o erro do tipo *II*, que é a falha em rejeitar a hipótese nula sendo esta falsa, é 0. Portanto, neste caso, as variações de acurácia são importantes.

Referências

[1] D. C. Montgomery and G. C. Runger, *Applied statistics and probability for engineers*, vol. 5. John Wiley; Sons, 2011.

[2] F. Campelo, "Lecture notes on design and analysis of experiments." https://github.com/fcampelo/Design-and-Analysis-of-Experiments, 2015.

[3] E. Walker and A. S. Nowacki, "Understanding equivalence and noninferiority testing." *Journal of general internal medicine*, vol. 26 2, pp. 192–6, 2011.