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.

O experimento deste trabalho foi desenvolvido no intuito de verificar as afirmações acima. Este consiste na comparação dos dois algoritmos citados baseando-se nos questionamentos abaixo:

Questionamentos

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

Características desejadas

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: $d^{\star}_{tempo} = 1.5$;
- Margem de não-inferioridade para acurácia: $\delta^{*}_{acuracia} = 0.05$.
- Potência desejada: $\pi = 0.8$;

Planejamento experimental

Caracterização do experimento

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.

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 convecional, onde a hipótese nula é de que o algoritmo proposto apresenta tempo de execução similar ao algoritmo atual. Já a hipótese alternativa é de que o algoritmo proposto é mais rápido que o padrão atual. Esta formulação é apresentada abaixo.

$$\begin{cases} H_0 : \mu_p - \mu_a = 0 \\ H_1 : \mu_p - \mu_a < 0 \end{cases}$$

- Referente ao questionamento 2 definiu-se o teste de não-inferioridade one a hipótese nula é que o algoritmo proposto apresenta acurácia inferior ao algoritmo atual. A hipótese alternativa é que o algoritmo proposto não apresenta acurácio inferior ao padrão atual como abaixo:

$$\begin{cases} H_0 : \mu_p - \mu_a = -\delta^*_{acurácia} \\ H_1 : \mu_p - \mu_a < 0 \end{cases}$$

Os valores de μ_p e μ_a são as médias de tempo de execução dos algoritmos *proposto* e *padrão atual* respectivamente.

Tamanho amostral

Afim de se atingir as características exigidas para o experimento é necessário definir o tamanho amostral a ser utilizado nos testes.

No caso de amostras pareadas, entende-se por amostras, o números de instâncias conforme exposto por [3], ou problemas a serem resolvidos pelos algoritmos analisados neste trabalho. Outra consideração é o número de repetições a serem adotadas por instância. Para este estudo adotou-se o valor de 30 repetições por instância seguindo recomendação expressas em [2] de que um valor heurístico para esta finalidade deve ser maior ou igual a 30.

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.

Como não há informação histórica da variância dos dados utilizou-se a equação abaixo como uma primeira aproximação para tamanho amostral do *Estudo Piloto*.

$$n_{pilot} \approx 2 \left(\frac{z_{\alpha_n/2}}{e_n} \right)^2$$

A equipe adotou um erro relativo $e_n = 0.10$, o que resultou em um tamanho amostral de 800. O autor [2] sugere cautela no uso da equação acima, sendo que está pode fornecer um número muito maior do que o número de amostras necessárias para um *Estudo Piloto*.

Contudo, o valor acima foi utilizado como estimativa inicial para iterações na equação abaixo. A equipe adotou amostras de mesmo tamanho e que se dispõe do menor tamanho de interesse prático para os ganhos de tempo $d^{\star}_{tempo}$.

$$n = 2(\frac{1}{d^{\star}_{tempo}})^2(t_{\alpha/2} + t_{\beta})^2$$

,

Os valores de $t_{\alpha/2}$ e t_{β} foram subtituidos para $z_{\alpha/2}$ e z_{β} na equação acima considerando-se a prmeira iteração com 800 amostras. As iterações convergiram rapidamente para um valores de 17 amostras a serem utilizadas no *Estudo Piloto* e foram geradas conforme abaixo:

- Nascimento: 21/11/1992;
- Número de instâncias de cada problema: 17;
- Número de execuções: 30;

```
#dados <- read.csv("1992-11-21_17_30.csv", header = T)
#head(dados)
```

Observou-se que os dados gerados continham inconsistências por apresentarem valores negativos para o tempo de execução do algoritmos como se pode observar abaixo.

Inserir código para verificar dado negativo

neg = dados$Time.s < 0
*neg*2 = *dados*Accuracy < 0 & dados$Accuracy > 1\ dados[neg,]

Os dados negativos de tempos de execução e respectivas acurácias foram retirados para se evitar contaminação do resultados. Para tal, considerou-se que a retida de 2 valores não seria representativa na massa amostral adotada.

Inserir código substituindo valores negativos por NA

dados[neg,]$Time.s = NA$
dados[*neg*,]Accuracy = NA \ dados[neg,]

Por meio dos dados do *Estudo Piloto* obteve-se o desvio padrão das amostras dos algoritmos. Estes valores foram utilizados para se obter o tamanha amostral adequado às caracterísitcas exigidas para os testes. Utilizando-se a função "calcN_tost2" fornecida por [2], foram verificados os tamanhos amostrais mínimos necessários considerando-se o tempo de execução e acurária dos algoritmos.

```
v_nTime = calcN_tost2(alpha = 0.05,
            beta = 0.2,
            diff_mu = 1,
            tolmargin = min(v_dataSD$Time.s),
            s1 = v_dataSD$Time.s[1],
            s2 = v_dataSD$Time.s[2]
            )

v_nAcc = calcN_tost2(alpha = 0.05,
                     beta = 0.2,
```

```
                        diff_mu = 0.01,
                        tolmargin = 0.05,
                        s1 = v_dataSD$Accuracy[1],
                        s2 = v_dataSD$Accuracy[2]
                       )
```

Para a geração e coleta dos dados dos foram adotados os valores de tamanho amostral verificado para o tempo de execução, sendo este o valor que atende às características exigidas para os testes com relação ao tempo de execução e acurácia. Os dados foram gerados conforme abaixo:

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

As amostras geradas para os testes também apresentaram tempo de execução negativo e estas foram tratadas conforme as amostras do *Estudo de Piloto*.

Inserir código para verificar dado negativo

neg = dados$Time.s < 0$
$neg2 = dados$Accuracy < 0 & dados$Accuracy > 1\ dados[neg,]

Inserir código substituindo valores negativos por NA

dados[neg,]$Time.s = NA$
$dados[neg,]$Accuracy = NA \ dados[neg,]

Teste das Hipóteses

Resultados

O teste de hipótese referente ao questionamento 1 foi implementado como abaixo e seu resultado indica que se deve rejeitar a hipótese nula.

```
# hypothesis test Time

v_diffTime = subset(v_data, Algorithm=='Proposed')$Time.s - subset(v_data, Algorithm=='Standard')$Time.s
v_tTestTime = t.test(v_diffTime,
                     conf.level = 0.05,
                     mu=0,
                     alternative = 'less'
                     )
v_pTime = v_tTestTime$p.value
```

O gráfico bloxplot abaixo evidência uma diferença entre as médias dos tempos de execução. Podemos ver que o algoritmo proposto possui a média abaixo em relação ao tempo médio de execução do algoritmo padrão.

Inserir chunk

boxplot(Time.s ~ Algorithm, data = mean.Time, col = (c("green", "blue")), main = "Tempo - Média das Instâncias", names = c("Proposed", "Standard"), xlab = "Algoritmo", ylab = "Tempo de Execução")

O teste de hipótese referente ao questionamento 2 foi implementado como abaixo e seu resultado indica que também se deve rejeitar a hipótese nula.

```
# hypothesis test acc
v_diffAcc = subset(v_data, Algorithm=='Proposed')$Accuracy - subset(v_data, Algorithm=='Standard')$Accu
v_tTestAcc = t.test(v_diffAcc,
                    mu = -0.05,
                    conf.level = 0.05,
                    alternative = 'greater'
                    )
v_pAcc = v_tTestAcc$p.value
```

Refazer o texto abaixo

O gráfico bloxplot abaixo evidência uma diferença entre as médias dos tempos de execução. Podemos ver que o algoritmo proposto possui a média abaixo em relação ao tempo médio de execução do algoritmo padrão.

Inserir chunk

boxplot(Accuracy ~ Algorithm, data = mean.Time, col = (c("green", "blue")), main = "Tempo - Média das Instâncias", names = c("Proposed", "Standard"), xlab = "Algoritmo", ylab = "Tempo de Execução")

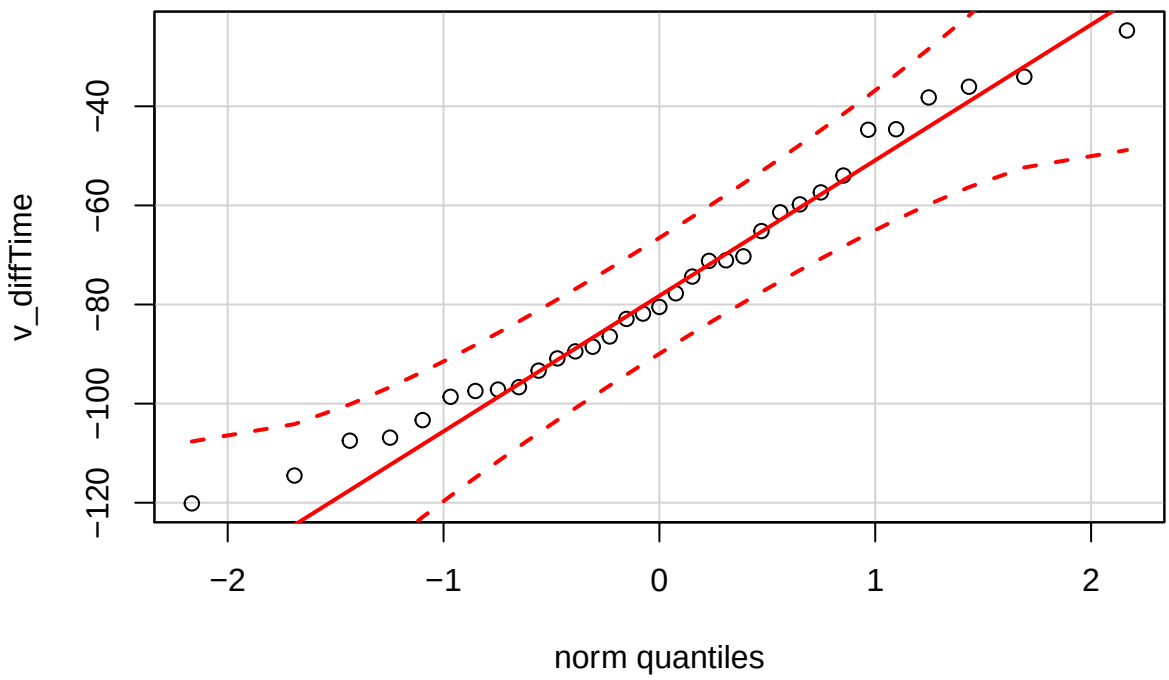
Validação dos testes

Como não se tem informações sobre a variância da população, o grupo adotou o teste de t student, assumindo a premissa de normalidade e independência.

A premissa de normalidade para o teste do questionamento 1 foi testada inicialmente de forma visual a partir do gráfico QQplot

```
## Hypothesis validation - Time

# Normality
qqPlot(v_diffTime)
```



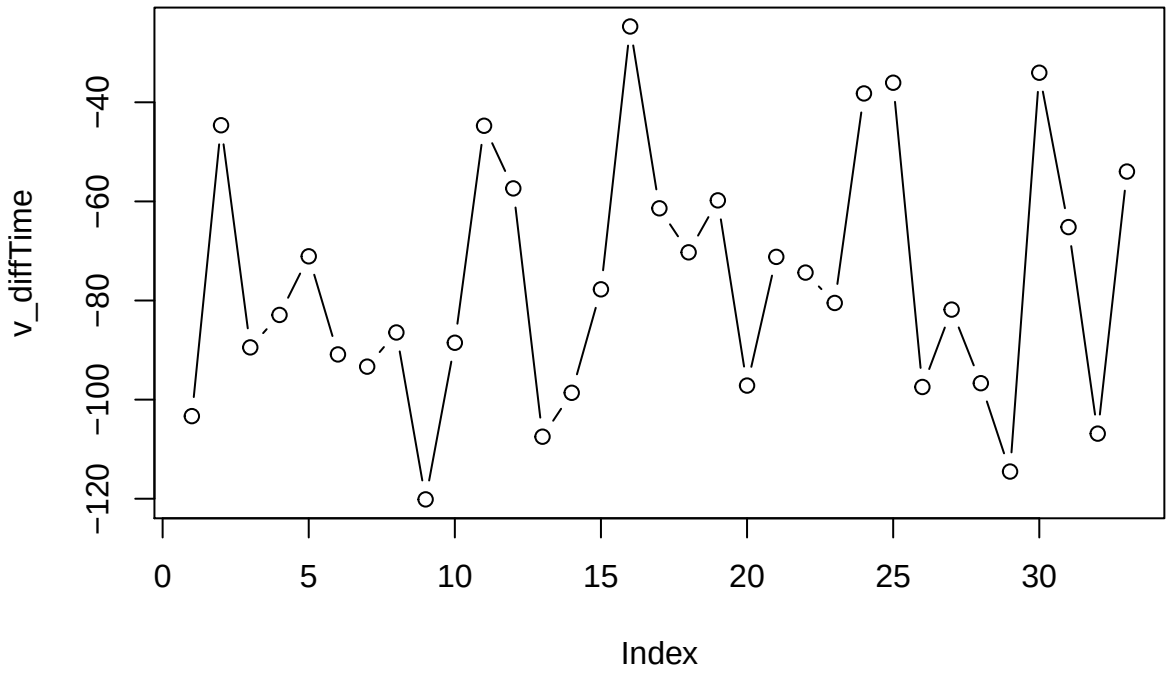
O gráfico QQplot para o tempo de execução sugere normalidade. Sendo assim foi aplicado o teste de Shapiro-Wilk para normalidade

```
v_shapiroTime = shapiro.test(v_diffTime)
```

O resultado sugere normalidade considerando a diferença de tempo médio de execução dos algoritmos analisados neste estudo de caso.

Outro teste utilizado foi o de verificação de independência de Durbin-Watson:

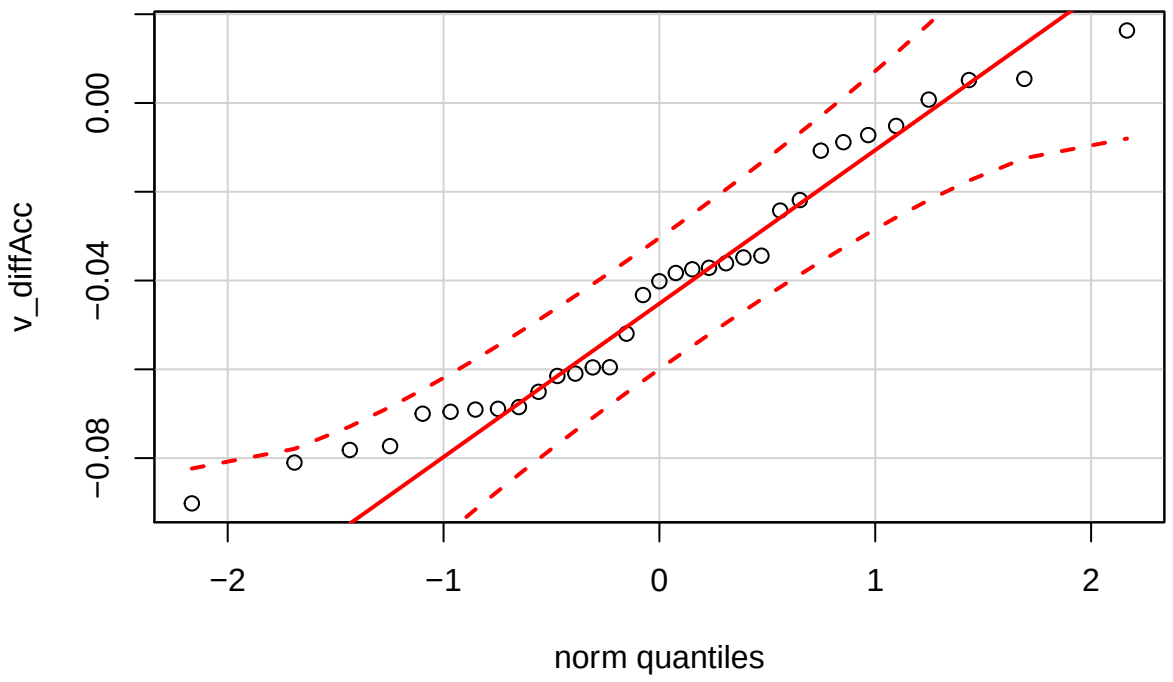
```
# Independence
v_dwTime = dwtest(v_diffTime~1)
plot(v_diffTime, type='b')
```



A premissa de normalidade para o teste do questionamento 2 foi testada inicialmente de forma visual a partir do gráfico QQplot.

```
## Hypothesis validation - Acc

# Normality
qqPlot(v_diffAcc)
```



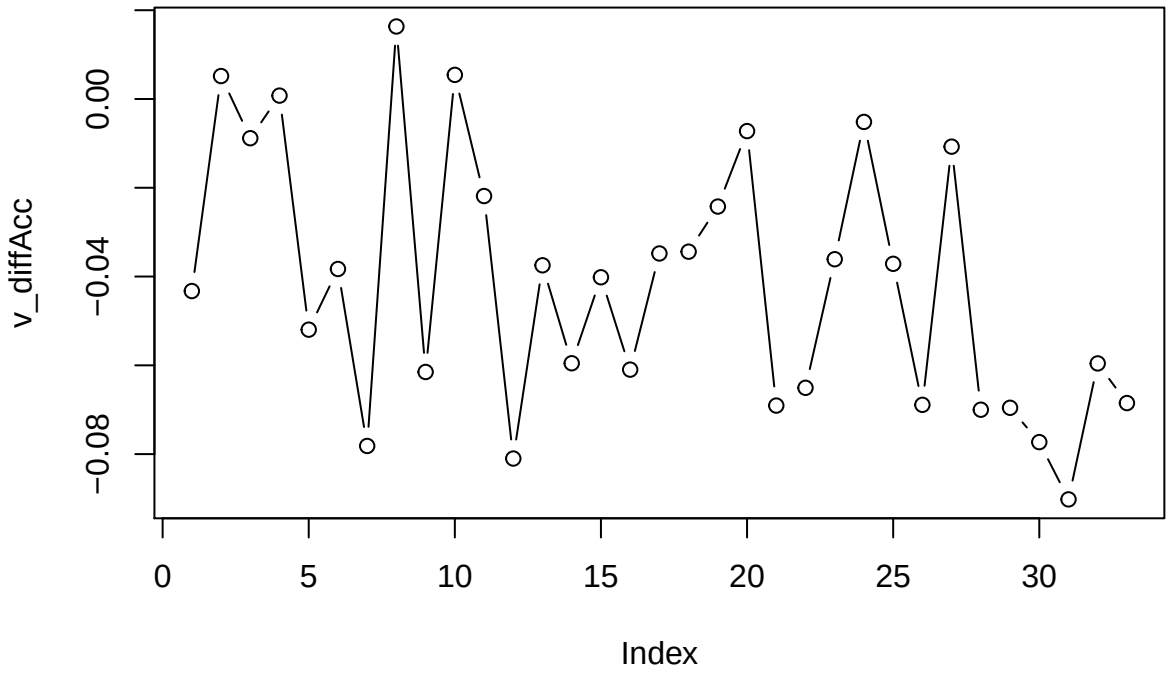
O gráfico QQplot para a acurácia sugere normalidade. Sendo assim foi aplicado o teste de Shapiro-Wilk para normalidade.

```
v_shapiroAcc = shapiro.test(v_diffAcc)
```

O resultado sugere normalidade considerando a diferença de acurácia média dos algoritmos analisados neste estudo de caso.

Também se verificou a independência por meio do teste Durbin-Watson para a acurácio dos algoritmos deste estudo:

```
# Independence
v_dwAcc = dwtest(v_diffAcc~1)
plot(v_diffAcc, type='b')
```



Conclusão

POR FAZER

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.