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

1. Descrição do Problema

O objetivo do experimento proposto é avaliar uma nova técnica proposta para simplificação de modelos em algoritmos de classificação, baseada em inferência estatística. Dessa forma, será realizada a comparação do algoritmo de classificação original e do método simplificado proposto. Para realizar o estudo, ambos serão executados em bases de dados da literatura.

Segundo os pesquisadores responsáveis pelo novo algoritmo proposto, este apresenta melhoria significativa em relação ao algoritmo original no tempo de execucão e não resultar 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: $d_t^\star = 1.0$;
- Margem de não-inferioridade para acurácia: $\delta_{acc}^* = 0.05$;
- Potência desejada: $\pi = 0.8$.

2. Planejamento Experimental

Os dados experimentais utilizados foram obtidos através de simulação, por meio de um aplicativo web. Os dados gerados informam o tempo necessário para a classificação (*Time.s*) e acurácia (*Accuracy*) de cada um dos algoritmos em cada instância e em cada execução. A data de nascimento do membro mais jovem da equipe (21/11/1992) é parâmetro utilizado como semente do gerador de números da simulação. O número de instâncias utilizadas e de execuções por instâncias deve ser selecionado no aplicativo.

O experimento envolve comparações entre métodos aplicados em diferentes instâncias de problemas de classificação. Conforme apresentado em [1], a variabilidade decorrente das características de cada problema de teste é uma forte fonte de variação espúria quando não considerada na análise dos resultados. Para eliminar a influência dessas variações, deve-se realizar o pareamento das medições por instância de teste.

Para isso, pode ser aplicado o teste de hipótese t pareado [2], comparando as médias de ambos os algoritmos. No entanto, uma segunda alternativa é realizar o teste sobre as diferenças das médias dos algoritmos, de maneira que o efeito das instâncias se cancela. Dessa forma, define-se para cada par j das n observações de médias (μ_{1j}, μ_{2j}) a diferença $d_j = \mu_{2j} - \mu_{1j}, \forall j \in (1, ..., n)$.

2.1 Definição de Hipóteses

Para cada uma das questões levantadas no experimento, foi estabelecido um par de hipóteses.

2.1.1 Ganhos de Tempo

Definiu-se o teste de hipótese unilateral convencional. A hipótese nula é de que a diferença do tempo de execução médio do algoritmo proposto e do algoritmo original é nula, enquanto a hipótese alternativa estabelece que o algoritmo proposto é mais rápido que o original (diferença menor que zero). Esta formulação é apresentada abaixo, na qual μ_p^t representa a média de tempo do método proposto e μ_o^t a média de tempo do algoritmo original.

$$\begin{cases} H_0 : \mu_p^t - \mu_o^t = 0 \\ H_1 : \mu_p^t - \mu_o^t < 0 \end{cases}$$

2.1.2 Não-inferioridade da Acurácia

Foi definido um teste de não-inferioridade do algoritmo proposto em relação ao atual. A hipótese nula estabelece que a diferença da acurácia média do algoritmo proposto e do algoritmo orignal é maior que a m argem de não-inferioridade estabelecida (δ_{acc}^*), enquanto a enquanto a hipótese alternativa propõe que a diferença das acurácias é menor que essa margem. Sendo μ_p^a a média da acurácia do algoritmo proposto e μ_o^a a média da acurácia do algoritmo original, tem-se:

$$\begin{cases} H_0 : \mu_p^a - \mu_o^a = -\delta_{acc}^* \\ H_1 : \mu_p^a - \mu_o^a > -\delta_{acc}^* \end{cases}$$

2.2 Número de Execuções por Instância

Para problemas com variáveis pareadas, cada par de médias avaliadas em diferentes instâncias constitui uma amostra independente [3]. No entanto, é possível que, mesmo para observações coletadas sob condições homogêneas nas mesmas instâncias, exista o efeito de perturbações aleatórias [2]. Uma forma de reduzir esse efeito é realizar repetidas execuções para cada instância e adotar como valor para a instância a média das diferentes execuções.

Já foi demonstrado que aumentar o número de instâncias testadas apresenta uma melhoria maior sobre a potência dos testes do que aumentar o número de execuções em cada instância. Apesar disso, se o custo de execução não é impeditivo, é recomendável realizar pelo menos 30 execuções em cada instância [1]. Esse valor foi adotado nesse trabalho, uma vez que o custo da simulação não é significativo.

2.3 Definição do Tamanho Amostral

Para a realizar o cálculo do tamanho amostral necessário para a potência estabelecida, é necessário conhecer o desvio padrão de cada variável observada. Uma vez que não há disponível nenhum conhecimento histórico sobre os processos em questão, deve ser realizado um estudo piloto para determinar os desvios.

Surge aí a necessidade de calcular o número de amostras para o estudo piloto. Inicialmente, utilizou-se a equação $n_{pilot} \approx 2\left(\frac{z_{\alpha_n/2}}{e_n}\right)^2$, onde e_n representa o máximo erro relativo permitido para o tamanho da amostra. Estabelecendo $e_n = 0.1$, obtém-se $n_{pilot} = 800$. No entanto, essa equação pode resultar em tamanhos de estudo piloto maiores que o tamanho amostral necessário para o experimento [1].

Para contornar esse problema, um caminho alternativo foi tomado. O tamanho da amostra pode ser calculado pela equação abaixo:

$$n = 2\left(\frac{\hat{\sigma}}{\delta^\star}\right)^2 (t_{\alpha/2} + t_\beta)^2$$

Uma vez que, para o experimento do tempo o valor de $d_t^\star = \frac{\delta^\star}{\hat{\sigma}}$ é conhecido, é possível calcular o tamanho amostral mínimo necessário com a equação:

$$n = 2\left(\frac{1}{d_t^\star}\right)^2 (t_{\alpha/2} + t_\beta)^2$$

Nessa fórmula, os valores $t_{\alpha/2}$ e t_β são dependentes de n. Para solucionar esse problema, eles são substituídos por $z_{\alpha/2}$ e z_β e a equação é testada iterativamente até convergência (implementação em anexo no arquivo *calcN.R*). Dessa forma, foi encontrado o valor $n = 17$. Esse valor é significativamente menor que o valor $n_{pilot} = 800$ obtido anteriormente, o que é indicativo de que esse valor era superestimado.

Com base nos resultados alcançados, foi realizado um estudo piloto com 17 amostras. A partir desse estudo, foram determinados os desvios padrão de tempo e acurácia para o algoritmo proposto e original:

- $sd_p^t = 17.5964261$
- $sd_o^t = 19.6324136$
- $sd_p^a = 0.0198$
- $sd_o^a = 0.0237973$

A partir desses valores, utilizou-se a fórmula abaixo (implementada no arquivo *calcN_tost2.R* [1]) para calcular o tamanho amostral mínimo para os experimentos de tempo e acurácia:

$$n >= (t_{\alpha;\nu} + t_{(1-c)\beta;\nu})^2 \left(\frac{\hat{\sigma}_1^2 + \hat{\sigma}_2^2}{\delta^\star - \Delta\mu^\star}\right)^2$$

Utilizando essa equação, determinou-se um tamanho amostral mínimo de $n = 33$ amostras para o teste de tempo e $n = 5$ amostras para o experimento com acurácia. O valor definitivo de amostras utilizado na coleta de dados foi então determinado como o máximo dos dois, 33.

2.4 Tratamento e Validação dos Dados

Considerando o experimento realizado, foi criada uma rotina para validação dos dados obtidos e identificação de erros. Para cada execução do algoritmo de classificação, as seguintes condições devem se aplicar:

1. Tempo de execução > 0
2. Acurácia $\in [0, 1]$

Caso os valores de uma execução não atendam essas condições, ela é descartada.

3. Análise Estatística

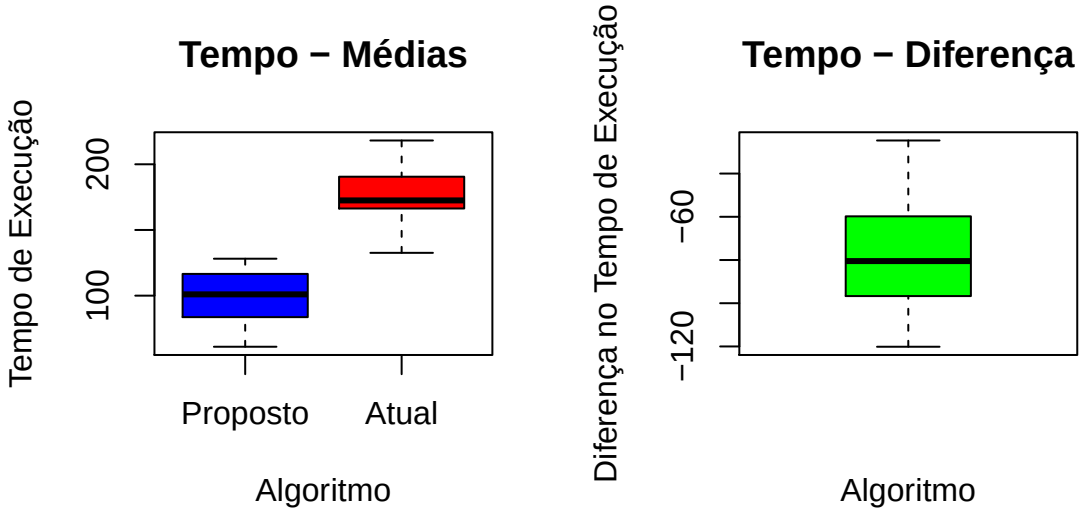
3.1 Teste de Hipóteses

3.1.1 Ganho de Tempo

Realizando o teste de hipóteses apresentado na Seção 2.1.1, obtém-se $p = 2.6192257 \times 10^{-18}$. Dessa forma, é possível rejeitar a hipótese nula com um nível de confiança de 95% e aceitar a hipótese a hipótese alternativa

que estabelece que o novo método proposto tem ganhos de tempo em relação ao método original, considerando um tamanho de efeito $d^{\star} = 1$.

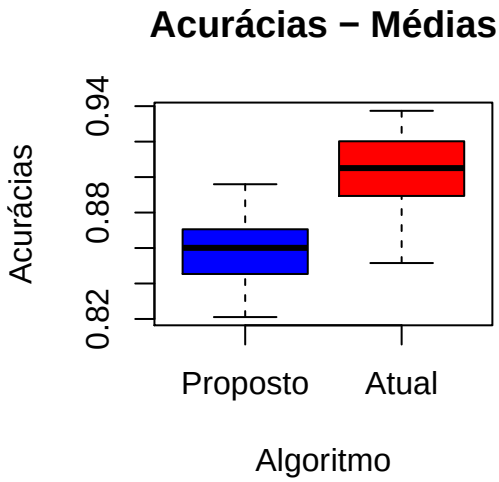
Os bloxplots abaixo evidenciam a diferença entre as médias dos tempos de execução. Nota-se valores muito menores para o tempo de execução do algoritmo proposto. No boxplot da diferença, pode-se observar que a média é menor que zero, o que indica que o algoritmo proposto possui média de tempo de execução menor que a média do algoritmo original.

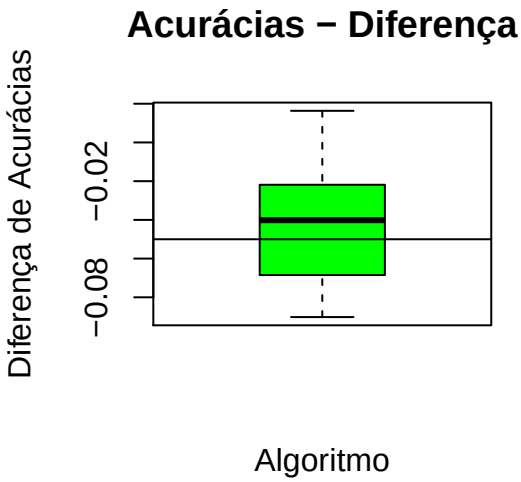


3.1.2 Não-inferioridade da Acurácia

O teste de hipóteses proposto na Seção 2.1.2 apresenta $p = 0.0611063$. Esse resultado não permite rejeitar a hipótese nula com um nível de confiança de 95%, considerando a margem de não-inferioridade $\delta^{*}_{acc} = 0.05$ estabelecida. Dessa forma, não é possível afirmar que o algoritmo proposto não apresenta acurácia inferior ao original.

O gráfico bloxplot referente ao experimento 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. Portanto não foi possível estabelecer a não-inferioridade do algoritmo em relação à sua acurácia.



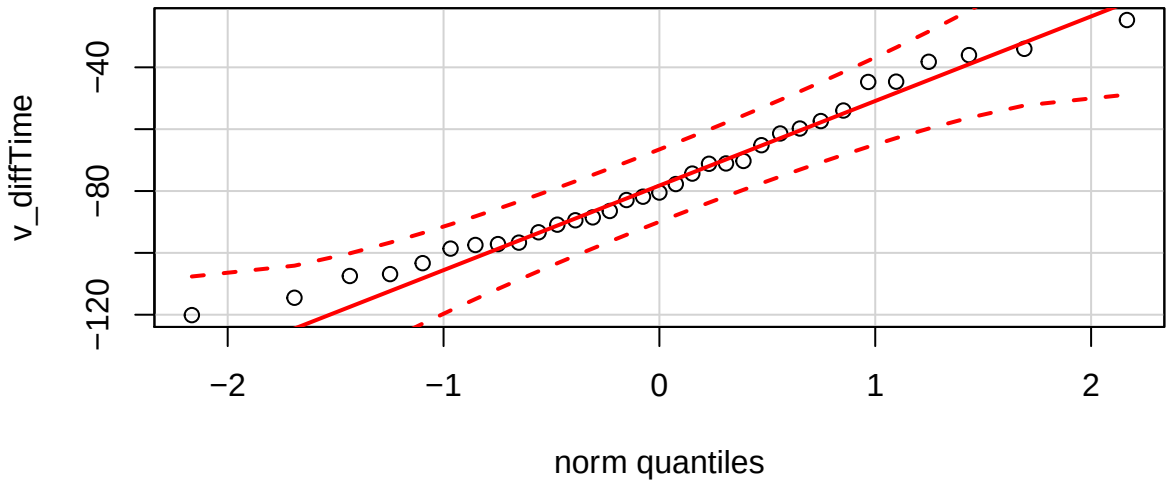


3.2 Validação das Premissas

3.2.1 Tempo

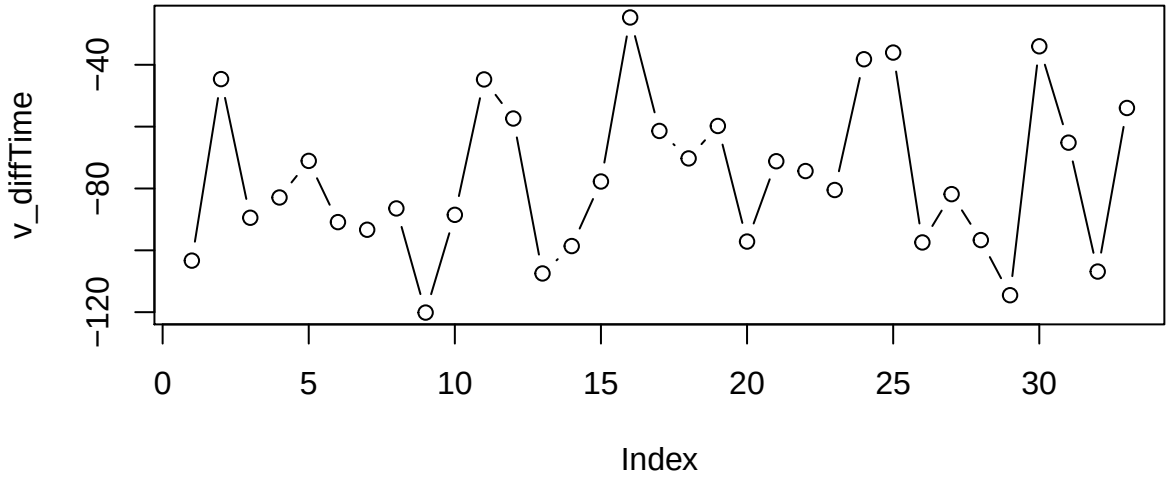
Normalidade

O QQPlot das diferenças de tempo entre os algoritmos é indicativo da normalidade de sua distribuição. Para confirmar esse resultado, é realizado o teste de Shapiro-Wilk. O teste apresenta $p = 0.5151005$, de maneira que não é possível refutar a hipótese nula de que os dados apresentam distribuição normal.



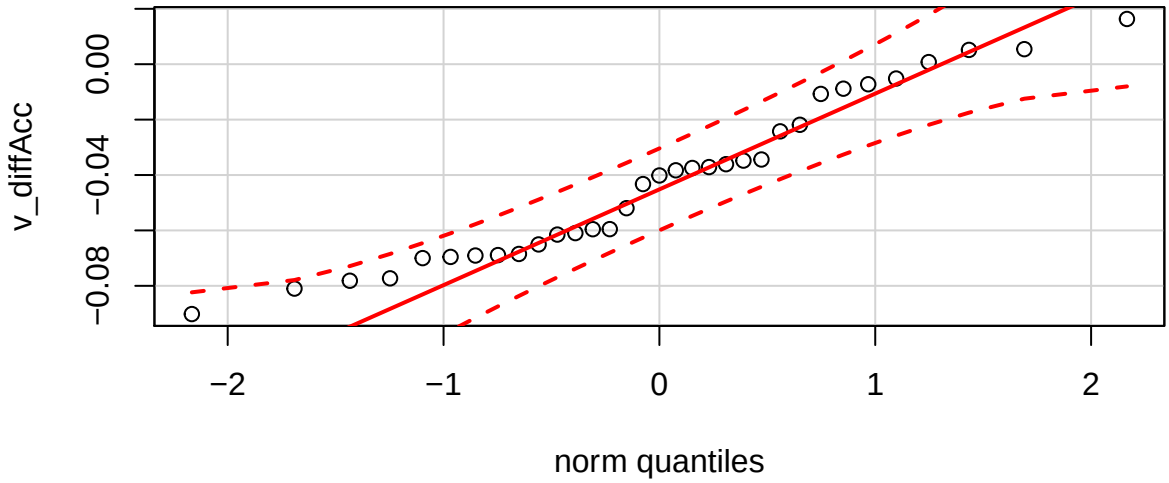
Independência

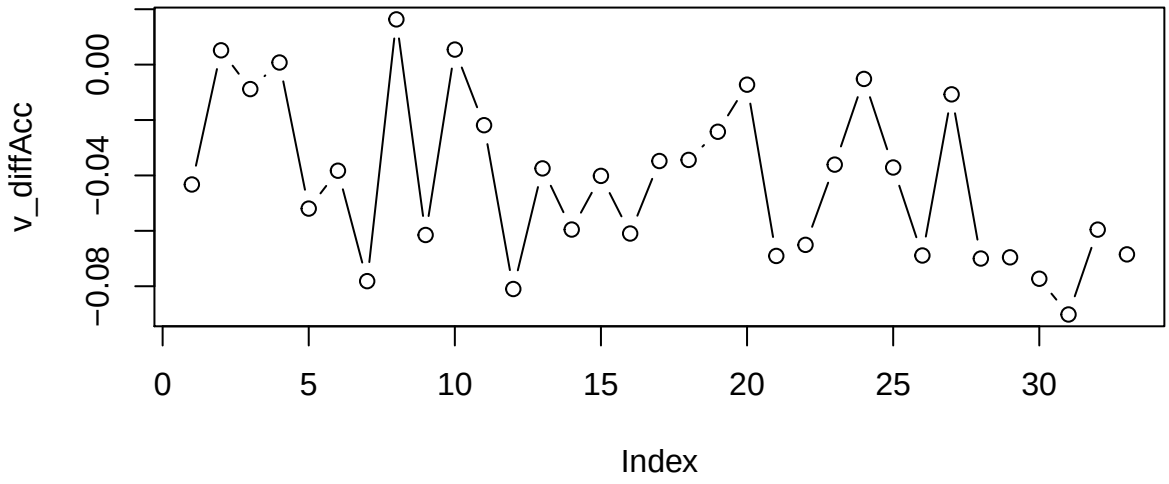
O plot dos valores ordenados de diferenças de tempo entre os algoritmos não apresenta nenhum indício de dependência temporal dos valores. O teste de autocorrelação serial Durbin-Wastson apresenta $p = 0.4003581$, o que reforça a hipótese de que não há autocorrelação serial entre as amostras.



Acurácia

Os mesmos procedimentos são realizados para a acurácia. O teste de normalidade Shapiro-Wilk apresenta $p = 0.121192$, enquanto o teste de autocorrelação serial Durbin-Watson apresenta $p = 0.3452157$. Nenhum dos resultados permite refutar as premissas de normalidade e independência.





4. Discussão e Conclusões

Os testes de hipótese realizados levam as seguintes conclusões:

1. É possível rejeitar a hipótese de que os algoritmos possuem tempos de execução equivalentes com grau de confiança de 95% para um tamanho de efeito $d^* = 1$. Esse resultado indica que o novo algoritmo apresenta tempos de execução melhores.
2. Não é possível refutar a hipótese de que a diferença de acurácia entre os algoritmos é maior que o tamanho de efeito $\delta^* = 0.05$ de 95%. Esse resultado não permite afirmar não inferioridade da acurácia do algoritmo proposto.

Vale notar que o teste de acurácia apresenta potência 1. Dessa forma, é seguro afirmar que não é possível estabelecer sua não inferioridade.

Os resultados indicam, portato, um trade-off entre os métodos avaliados. Enquanto o método proposto apresenta ganhos de tempo de execução de pelo menos um desvio padrão em relação ao anterior, não é possível garantir sua não inferioridade em relação a acurácia. A análise descritiva sugere que o método original possui maior acurácia. Assim, a utilização de cada método depende dos requisitos da aplicação. Se o tempo de execução for prioridade e uma ligeira perda na acurácia for aceitável, recomenda-se o método proposto. Se a acurácia da classificação é prioridade e tempos maiores são aceitáveis, recomenda-se o método original.

Referências

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

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

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