X Simpósio Brasileiro de Melhoramento Animal
Uberaba, MG – 18 a 23 de agosto de 2013

GS3-CV : um aplicativo multiplataforma para validação de predições genômicas

Katia C. Lage dos Santos[1], Wagner Arbex[1], Daniel J. De Abreu Santos [2], Adam H. Utsunomiya[3], Fernanda N. Almeida[4], Marcos Vinícius G. B. da Silva[5]

[1]Analista A - Embrapa Gado de Leite, Juiz de Fora.. e-mail: {katia.santos, wagner.arbex}@embrapa.br
[2] Pós Graduando - FCAV/UNESP/Jaboticabal. e-mail: daniel_jordan2008@hotmail.com
[3] Pós Graduando - FCAV/UNESP/Jaboticabal. e-mail adamtaiti@gmail.com
[4] Bolsista Fapemig - Embrapa Gado de Leite, Juiz de Fora. e-mail almeida.fn@gmail.com
[5] Pesquisador A - Embrapa Gado de Leite, Juiz de Fora. e-mail marcos.vb.silva@embrapa.br

Resumo: As ferramentas computacionais disponíveis para aplicação dos modelos de Seleção Genômica Ampla visam estimar os valores genômicos dos animais de uma população. Na prática, os resultados obtidos viabilizam a seleção precoce de indivíduos para características de importância econômica. Com o objetivo de definir uma metodologia que aumente a precisão nos resultados obtidos, foi implementado um aplicativo multiplataforma, denominado GS3-CV, que utiliza a técnica de validação cruzada. O GS3-CV é uma extensão do software GS3, bem reconhecido pela área, o qual permite o uso de métodos relevantes para predição do valor genético genômico. Os experimentos iniciais realizados mostram que o uso da validação cruzada pela metodologia proposta produz resultados mais próximos do esperado, comparado à execução clássica do GS3.
Palavras–chave: distribuição de valores genéticos, GS3, Seleção Genômica Ampla, validação cruzada

GS3-CV: a multi-platform application to genome prediction validation

Abstract: The computational tools available to apply the Genome-wide Selection intend to estimate the genomic breeding values of individuals in a population. In practice, genomic selection allows breeders to identify genetically superior animals at a much earlier age for economically important traits. Aiming to define a methodology that increases the precision of the results obtained, it was implemented a multi-platform application, named GS3-CV, which uses the cross-validation technique. The GS3-CV is an extension of the software GS3, well known in the area because it implements methods to predict the genomic breeding value. The results obtained during initial trials are consistent and permit to infer that the use of cross-validation produces more consistent solutions for marker effects.
Keywords: *cross-validation, genetic value distribution, Genome Wide Selection, GS3*

Introdução

A eficiência do melhoramento genético depende da identificação de genótipos superiores pelos melhoristas. Nesse sentido, a seleção desempenha papel fundamental na definição dos acasalamentos a serem realizados, visando a obtenção de novos genótipos e na indicação dos indivíduos superiores a serem usados comercialmente (RESENDE et al, 2008).

Nesse contexto, as ferramentas de Seleção Genômica Ampla (*Genome-wide Selection* – GWS) viabilizam a predição dos valores genético-genômicos (GBVs) com base na leitura de marcadores genéticos e no uso de métodos preditivos. As predições geradas são comumente utilizadas para a seleção precoce de indivíduos no tocante à avaliação de características de importância econômica. Para atingir esse objetivo, existem alguns programas para estimar os GBVs, tais como o GS3, GEBV e GenSel. As diferenças entre esses programas residem nos modelos disponíveis. Neste trabalho, foi usado o GS3, pois os modelos implementados são de uso relevante para pesquisas em genômica e melhoramento genético.

Segundo a literatura relacionada, um dos principais desafios da GWS é a estimação de um grande número de efeitos a partir de um número limitado de observações. Aliando essa afirmativa ao fato do GS3 obter as estimativas de um modelo sobre todo o conjunto de dados, nesse trabalho foi realizada uma medida de extensão do referido software. Com o intuito de obter uma distribuição empírica e possivelmente próxima da normalidade dos valores genéticos, foi desenvolvido o aplicativo GS3 Cross-validation (GS3-CV) que implementa a técnica de validação cruzada. Seu princípio básico consiste em treinar um modelo de GWS em um conjunto de dados, denominado população de treinamento ou teste, na qual são verificados os marcadores que explicam os locos que controlam as características, bem como

são estimados os seus efeitos. Após esse passo, é avaliada a adequação dos resultados obtidos em uma população distinta, chamada população de validação, a qual, por não ter sido envolvida na predição dos efeitos dos marcadores, possui independência entre os erros dos valores genéticos genômicos e dos valores fenotípicos. A correlação entre esses valores é predominantemente de natureza genética e equivale à capacidade preditiva da GWS em estimar os valores genômicos. Com isso, as análises decorrentes do método empregado podem ser generalizadas para os diferentes grupos de dados.

Material e Métodos

O GS3 foi desenvolvido para avaliação genômica ampla, podendo utilizar os modelos BLUP Genômico (ou G-BLUP), Bayesian Lasso e Bayes Cpi. Distribuído sob a licença GNU, são disponibilizados o código-fonte e os arquivos executáveis para os sistemas Windows e Linux. Deve ser executado por linha de comando, informando o caminho do arquivo de parâmetros.

O aplicativo GS3-CV, aqui apresentado, é uma extensão do GS3. Ele foi desenvolvido em Perl, viabilizando a sua execução em diferentes plataformas (Windows, Linux etc.). A única pré-condição para seu funcionamento é a instalação do interpretador da linguagem na plataforma de execução.

Para a execução do GS3-CV buscou-se mimetizar o processo já utilizado no GS3. Assim, por linha de comando são informados quatro parâmetros: o arquivo de parâmetros definido pelo GS3; o número de iterações do GS3-CV; o número de observações a serem retiradas dos arquivos de dados e genótipos; e um *flag* binário (0-não; 1-sim) indicando a criação dos arquivos com as observações excluídas.

Em cada iteração do GS3-CV, o primeiro passo consiste em gerar dois subconjuntos disjuntos de dados. Esse cenário viabiliza a realização da validação cruzada (*cross-validation*). Este é um método estatístico para estimativa de parâmetros de um modelo, fase esta denominada de treinamento. Ele realiza a divisão dos dados em dois grupos: um para aprendizado ou treinamento e o outro para validação do mesmo modelo. A forma básica dessa técnica é denominada *k-fold cross-validation*. Inicialmente, o conjunto de dados é particionado em *k* segmentos de mesmo tamanho. Em seguida, *k* iterações de treinamento e validação são realizadas de forma que, em cada iteração, um subconjunto diferente de dados é manipulado para validação, enquanto que os demais *k*-1 subconjuntos restantes são utilizados para treinamento. A Figura 1 demonstra um exemplo com *k*=3. As seções em cinza claro são utilizadas para treinamento, enquanto que as em cinza escuro são utilizada para validação.

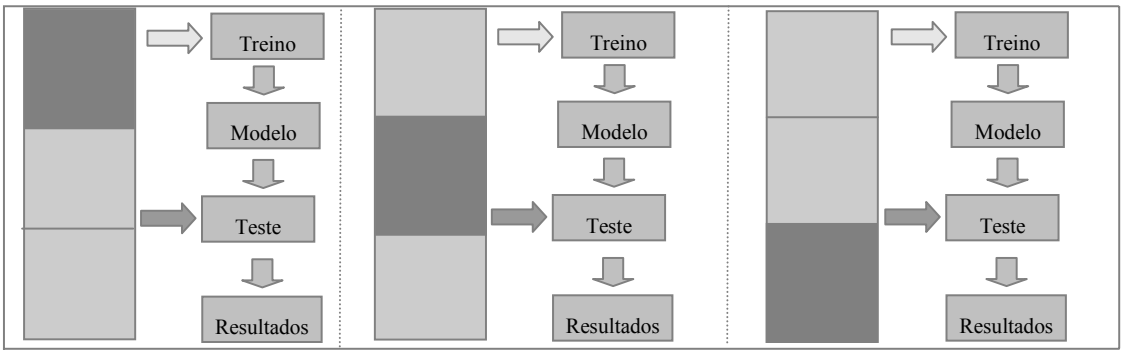

Figura 1. Esquema do funcionamento da validação cruzada em três iterações (3-*fold cross validation*)

Diferentemente das versões clássicas da validação cruzada, no GS3-CV a seleção das observações contidas nas populações de teste e validação não é feita de forma estática para todas as iterações. De acordo com o número de observações a serem excluídas, em cada iteração do GS3-CV são selecionadas aleatoriamente as entradas dos arquivos de dados e de genótipo do conjunto de treinamento. Para isso, foram utilizadas duas funções nativas da biblioteca Perl: *rand*, que gera os números pseudo-aleatórios, e *srand*, que ajusta a semente *de rand* com o relógio da máquina. A vantagem dessa estratégia é o melhor aproveitamento da base de dados, uma vez que uma mesma observação pode participar do conjunto de treinamento em uma iteração, mas em outra pode participar da população de validação.

O passo anterior é repetido para cada uma das iterações do GS3-CV. Ao término de todas as iterações, o aplicativo realiza o cálculo do valor médio e do desvio-padrão dos valores de variância

genética. Além disso, é realizado o cálculo da média e do desvio-padrão da variância dos erros de predição (PEV) para cada animal em todas as rodadas. Esses valores calculados são utilizados para o cálculo da acurácia de cada animal, de acordo com a seguinte relação:

$$r(\hat{g},g) = \sqrt{\left(1 - \frac{PEV}{\sigma^2}\right)}$$

onde: ĝ é o valor genético
g é o valor verdadeiro

A acurácia é uma medida da correlação entre o valor genético estimado e os valores das fontes de informação. Com essa métrica é possível mensurar o quanto a estimativa que obtivemos é relacionada com o valor real do parâmetro e nos dá a confiabilidade daquela estimativa ou valor. Nesse trabalho, considerou-se que o valor genético do animal segue uma distribuição empírica dos valores genéticos, próxima da normalidade. Para o valor de σ^2, considerou-se a variância total estimada pelo GS3.

Resultados e Discussão

Os experimentos iniciais realizados utilizaram a base de dados e o arquivo de parâmetros (modelo Bayesian Lasso-VCE) disponibilizados com o código-fonte do GS3. Eles foram realizados em uma estação Ubuntu 12.04 LTS 64 bits, com processador de quatro núcleos de 2.27GHz e 4 GB de RAM. Inicialmente, foi realizada uma execução do GS3 para toda base de dados. Em seguida, foi utilizado o GS3-CV, considerando duas e dez iterações, e retirando-se 50 observações para os arquivos de validação. Para cada cenário, são apresentados na Tabela 1 os valores das médias e desvios-padrão das variâncias.

Tabela 1. Média e desvios-padrão das variâncias em três cenários

Métrica	GS3	GS3-CV (2 iterações)	GS3-CV (10 iterações)
Variância Aditiva (σ_a^2)	2,29e-04 (6,39e-05)	2,15e-04 (5,83e-05)	2,11e-04 (5,43e-05)
Variância de dominância (σ_d^2)	3,97e-06 (2,54e-06)	3,49e-06 (2,19e-06)	3,77e-06 (2,38e-06)
Variância genética (σ_g^2)	3,71 (0,32)	3,73 (0,33)	3,75 (0,32)
Variância de Ambiente Permanente (σ_p^2)	2,09 (0,23)	2,12 (0,23)	2,11 (0,23)
Variância total (σ_T^2)	0,86 (0,24)	0,81 (0,22)	0,80 (0,20)

Analisando os dados da Tabela 1, é possível constatar que o GS3-CV apresentou um menor valor para a variância total estimada, o que tende a representar uma maior aproximação dos valores genéticos preditos dos valores reais.

Sob o ponto de vista de processamento, o GS3-CV apresentou-se eficiente em termos de sua execução, obtendo, em seu *uptime*, o tempo de resposta de 2,26 segundos para a definição dos vinte conjuntos de dados e genótipos de treinamento e de validação. Tal observação sugere que a medida de complexidade de seu algoritmo não deve estabelecer a ordem de grandeza da complexidade de todo o procedimento, mesmo com a utilização de conjuntos de dados e/ou de genótipos da ordem de *gigabytes*.

Conclusões

Os experimentos mostraram que o uso da validação cruzada com o GS3-CV tende a produzir resultados mais próximos do esperado, comparado à execução clássica do GS3. Como trabalhos futuros, serão realizados o cálculo da acurácia para a população de validação com base nos valores preditos no treinamento e desenvolvimento de uma interface gráfica *desktop*.

Literatura citada

RESENDE, Marcos Deon Vilela, LOPES, Paulo Sávio Lopes, SILVA, Rogério Luíz e PIRES, Ismael Eleotério. Seleção genômica ampla (GWS) e maximização da eficiência do melhoramento genético. Pesquisa Florestal Brasileira, Colombo, n.56, p.63-77, jan./jun. 2008