GS3 Cross Validation (GS3-CV)

Manual de Uso

Considerações Iniciais

Este programa é um software livre; você pode redistribuí-lo e/ou modificá-lo sob os termos da Licença Pública Geral GNU, conforme publicada pela *Free Software Foundation*; tanto a versão 2 da Licença, como (a seu critério) qualquer versão posterior.

Este programa é distribuído na expectativa de que seja útil, porém, SEM NENHUMA GARANTIA; nem mesmo a garantia implícita de COMERCIABILIDADE OU ADEQUAÇÃO A UMA FINALIDADE ESPECÍFICA. Consulte a Licença Pública Geral do GNU para mais detalhes.

Você deve ter recebido uma cópia da Licença Pública Geral do GNU junto com este programa; se não, escreva para a Free Software Foundation, Inc., no endereço 59 Temple Street, Suite 330, Boston, MA 02111-1307 USA.

Apresentações

O GS3 Cross Validation (GS3-CV) é um aplicativo multiplataforma que implementa a técnica de validação cruzada para avaliar a habilidade de predição (ou acurácia) pelos modelos disponíveis no software GS3, tendo por base os valores genômicos dos animais (GBVs) de uma população. 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. O GS3 deve ser executado por linha de comando, informando o caminho do arquivo de parâmetros.

O GS3-CV foi desenvolvido em Perl, viabilizando a sua execução em diferentes plataformas. A única pré-condição para seu funcionamento é a instalação do interpretador da linguagem Perl na plataforma de execução.

Para a execução do GS3-CV devem ser informados quatro parâmetros na linha de comando: 1) o arquivo de parâmetros definido pelo GS3; 2) o número de iterações do GS3-CV; 3) o número de observações a serem retiradas dos arquivos de dados e genótipos; e 4) 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: um de treinamento e outro de teste. Esse cenário viabiliza a realização da validação cruzada

(cross-validation). Diferentemente das versões clássicas da validação cruzada, no GS3-CV, a seleção das observações contidas nas populações de teste e de validação não é feita de forma estática para todas as iterações. 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.

Ambiente Computacional

O GS3-CV foi desenvolvido em linguagem 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 Perl na plataforma de execução, caso esta não tenha o referido interpretador. Acesse o endereço eletrônico a seguir e selecione a versão do interpretador adequado para a plataforma: http://www.perl.org/get.html.

Instruções de Uso

Para executar o GS3-CV, em um terminal de linha de comando, deve ser invocado o interpretador Perl, informando o nome do arquivo do *script* correspondente ao GS3-CV e quatro parâmetros para a execução contínua do programa. Eles devem ser informados na seguinte ordem:

1. o arquivo de parâmetros definido pelo GS3;
2. o número de iterações do GS3-CV;
3. o número de observações a serem retiradas dos arquivos de dados e genótipos;
4. e um *flag* binário (0-não; 1-sim) indicando a criação dos arquivos com as observações excluídas.

Um exemplo de arquivo de parâmetros está apresentado a seguir:

```
DATAFILE
./exo_data.txt
PEDIGREE FILE
./pedigri.dat
```

```
GENOTYPE FILE
./exo_genotypes.txt
NUMBER OF LOCI (might be 0)
10946
METHOD (BLUP/MCMCBLUP/VCE/PREDICT)
VCE
SIMULATION
F
GIBBS SAMPLING PARAMETERS
NITER
10000
BURNIN
200
THIN
10
CONV_CRIT (MEANINGFUL IF BLUP)
1d-8
CORRECTION (to avoid numerical problems)
1000
VARIANCE COMPONENTS SAMPLES
var2
SOLUTION FILE
solutions2
TRAIT AND WEIGHT COLUMNS
1 0 #weight
NUMBER OF EFFECTS
5
POSITION IN DATA FILE TYPE OF EFFECT  NUMBER OF LEVELS
6 cross 1
5 add_animal 2272
7 perm_diagonal 2000
8 add_SNP 0
8 dom_SNP 0
VARIANCE COMPONENTS (fixed for any BLUP, starting values for VCE)
vara
2.52d-04 2
vard
1.75d-06 2
varg
3.56 2
varp
2.15 2
vare
0.19 2
RECORD ID
5
CONTINUATION (T/F)
F
MODEL (T/F for each effect)
T T T T T
A PRIORI a
1 1
a PRIORI D
1 1
USE MIXTURE
F
#OPTION BayesianLasso Tibshirani
#OPTION SNP_weights myfile 1
```

Uma chamada válida do GS3-CV via terminal está apresentado a seguir:

```
perl gs3_cv_v1.pl arquivoParametros.par 10 25 0
```

Na chamada anterior, o GS3-CV recebe como parâmetros: o arquivo do GS3 `arquivoParametros.par`; devem ser realizadas 10 iterações do mecanismo de validação cruzada; em cada iteração 25 animais farão parte da população de validação; os animais excluídos não serão armazenados em um arquivo à parte (*flag* com valor 0). Para cada iteração do GS3-CV é gerado um conjunto de arquivos referente às entradas do GS3 e à saída por este produzida. Para a chamada acima, onde foi informado um total de 10 (dez) iterações, os arquivos gerados pelo GS3-CV estão listados a seguir:

- 10 arquivos denominados `arquivoParametros.par_#`, onde `#` assume os valores de 1 até 10: em cada iteração do GS3-CV os arquivos de entrada e de saída informados no arquivo de parâmetros passado inicialmente recebem o número da iteração no final do seu nome;
- 10 arquivos denominados `exo_data.txt_#` e 10 arquivos denominados `exo_genotypes_#`, onde **#** assume os valores de 1 até 10: em cada iteração, o GS3-CV seleciona aleatoriamente 25 animais da população inicial (`exo_data.txt` e `exo_genotypes.txt`) que farão parte da população de validação. Os demais, são armazenados nos arquivos `exo_data.txt_#` e `exo_genotypes_#`, que correspondem respectivamente aos arquivos de dados e de genótipo da população de treinamento;
- 10 arquivos `solutions2_#` e 10 arquivos `var2_#` e 10 arquivos denominados `arquivoParametros.par_#_EBVs`, onde **#** assume os valores de 1 até 10: esses arquivos correspondem à saída do GS3 para cada conjunto de dados de entrada da iteração **#** correspondente;

Se for informado o valor 1 para o *flag* de criação de arquios de validação, dois outros tipos de arquivos serão gerados: 10 arquivos denominados `exo_data.txt_val_#` e 10 arquivos denominados `exo_genotypes_val_#`, onde **#** assume os valores de 1 até 10. Esses em arquivos correspondem, respectivamente, aos arquivos de dados e de genótipo da população de treinamento em cada iteração.