Implementação do Best Linear Unbiased Prediction (BLUP) em Python para avaliação genética animal

Caio Augusto C. Volpato[1]
Roberto Hiroshi Higa[2]

Resumo: Valores genéticos para animais em programas de melhoramento de grandes populações são obtidos pela resolução de Best Linear Unbiased Prediction (BLUP) correspondente ao modelo de avaliação utilizado. O sistema de equações lineares que resulta da resolução de BLUP contém centenas de milhares ou até milhões de equações para serem resolvidas. Neste trabalho, pretende-se avaliar a estratégia de representação desse sistema de equações lineares em memória Random Access Memory (RAM), baseada em matrizes esparsas, e sua resolução por meio do algoritmo de gradiente conjugado precondicionado (WATKINS, 2010), para avaliar sua aplicabilidade em grandes populações de animais. Até o presente momento, uma solução baseada em matrizes densas foi implementada e sua adaptação para utilizar matrizes esparsas encontra-se em curso. Este trabalho integra as atividades do Plano de Ação 4 do projeto "Desenvolvimento e implementação de metodologias genético-estatísticas em avaliações genéticas de gado de corte" (MENEZES, 2015) que pretende desenvolver um software Embrapa para avaliação genética de grandes populações de animais, livrando os programas de melhoramento genético coordenado pela Empresa Brasileira de Pesquisa Agropecuária (Embrapa) do pagamento de licenças de uso.

Palavras-chave: Melhoramento genético, grandes populações, BLUP, gradiente conjugado, Python.

[1] Estudante de Matemática Aplicada e Computacional da Unicamp, estagiário da Embrapa Informática Agropecuária, Campinas, SP.

[2] Engenheiro Eletricista, pesquisador da Embrapa Informática Agropecuária, Campinas, SP.

Introdução

Programas de melhoramento genético de animais utilizam medidas de características e de genealogia de dezenas de milhares, ou ainda milhões, de animas para avaliar a contribuição do fator genético na manifestação de características de interesse, em gerais econômicos, da cadeia pecuária (BOURDON, 2000; ROSA et al., 2013).

Para inferir a contribuição da genética na manifestação de características de interesse (valor genético) dos animais que participam de um programa de melhoramento utilizam-se modelos estatísticos conhecidos como modelos lineares mistos (BOURDON, 2000). Esses modelos acomodam tanto fatores fixos quanto aleatórios (JIANG, 2007). Nesses modelos, enquanto fatores como sexo e grupo de manejo são modelados como fatores fixos, o valor genético é modelado como um fator aleatório e todos eles são inferidos a partir da utilização dos dados observados para ajuste do modelo considerado (JIANG, 2007).

Uma das metodologias mais utilizadas para resolução de modelos lineares mistos é o BLUP, um estimador que minimiza a variância do erro de predição (JIANG, 2007). Na prática, a resolução do BLUP implica a solução de um sistema de equações lineares que, na prática, é obtida pela utilização de soluções computacionais (MRODE; THOMPSON, 2014).

No caso de programas de melhoramento genético de grandes populações de animais, a resolução do BLUP é ainda dificultada pelas dimensões do sistema de equações lineares que precisa ser resolvido (centenas de milhares ou milhões de equações) (MRODE; THOMPSON, 2014).

Em geral, a solução de sistemas de equações lineares de grande dimensão utiliza algoritmos iterativos, tais como Jacobi, Gauss-Seidel, sobre-relaxação sucessiva, gradiente conjugado e gradiente conjugado precondicionado (MISZTAL, 2011).

Em particular, no caso da resolução do BLUP aplicado a programas de melhoramento genético de grandes populações de animais, duas estratégias são possíveis. A primeira delas, bastante utilizada, consiste na modelagem da solução iterativa de sistemas de equações lineares com iteração nos dados (do inglês Interaction on Data - IOT), em que se mantém em memória apenas a solução corrente e para cada iteração do algoritmo lê-se o conjunto de dados de arquivos, adequadamente formatados (MISZTAL, 2011).

A segunda estratégia consiste em aproveitar-se do fato de que, o sistema de equações lineares para resolução do BLUP utiliza matrizes esparsas (matrizes esparsas são matrizes onde temos muitos elementos nulos, assim são armazenados apenas os elementos não nulos) para realizar toda a computação com o sistema de equações armazenado na memória RAM. Obviamente, essa estratégia é limitada pela quantidade de memória disponível e, a priori, não se presta a grandes populações de animais. A presente proposta de trabalho pretende explorar exatamente essa estratégia de solução do BLUP.

O objetivo do presente trabalho é: a) implementar a montagem do sistema de equações lineares em memória, utilizando estruturas de matrizes esparsas; b) implementar a resolução do BLUP em memória, utilizando o método do gradiente conjugado precondicionado (método iterativo para resolver sistemas lineares definidos positivos, estamos utilizando o método com suas matrizes precondicionadas, fazendo assim que o método convirja muito menos iterações); c) avaliar a escalabilidade dessa solução em termos de memória e tempo de processamento. Espera-se, dessa forma, obter informações sobre o tipo de população (dimensões) para as quais esse tipo de solução é aplicável, bem como a correspondente necessidade em termos de hardware (quantidade de memória requerida). Este trabalho integra as atividades do Plano de Ação 4 do projeto “Desenvolvimento e implementação de metodologias genético-estatísticas em avaliações genéticas de gado de corte” (MENEZES, 2015), que pretende desenvolver um software Embrapa para avaliação genética de grandes populações de animais, livrando os programas de melhoramento genético coordenado pela Embrapa do pagamento de licenças de uso.

Materiais e Métodos

Para realização da presente proposta de trabalho, a linguagem Python (PYTHON SOFTWARE FOUNDATION, 2015) será utilizada para integrar algoritmos e estruturas de dados disponíveis em diferentes bibliotecas, incluindo:

- SciPy: contém implementações do método do gradiente conjugado e matrizes esparsas.

- Numpy: contém funcionalidades para manipulações de vetores.
- PyPedal: contém a implementação de algoritmos necessários para a montagem do sistema de equações lineares para solução do BLUP, tais como a construção da matriz de parentesco (ou, equivalentemente, do inglês, numerator relationship matrix – NRM), e os algoritmos para obtenção dos elementos de sua inversa em que não seja necessário obter a matriz de parentesco, especificamente os métodos conhecidos como de Henderson e de Quaas.

Inicialmente, o pacote PyPedal (COLE, 2012) será utilizado diretamente na construção do sistema de equações lineares para solução do BLUP. Uma vez que ele não utiliza recursos de matrizes esparsas, num segundo passo, suas implementações da construção da matriz de parentesco e obtenção das matrizes inversas serão estudadas e adaptadas utilizando recursos de matrizes esparsas disponíveis na biblioteca SciPy.

Uma vez tendo construído a solução de BLUP, utilizando o algoritmo de gradiente conjugado precondicionado com o sistema representado em memória RAM por meio de matrizes esparsas, ele será resolvido considerando diferentes tamanhos de populações, sendo medidos o consumo de memória RAM e o tempo de processamento.

Para esse trabalho, será considerado o modelo animal simples, ilustrado na equação matricial $y = Xb + Za + e$, onde y é o vetor com as observações, X é matriz de design dos efeitos fixos, Z é matriz de design dos efeitos aleatórios, a é o vetor com os efeitos animais aleatórios, b é o vetor com os efeitos fixos e é o vetor com os efeitos residuais extensamente explorado em exemplos em Mrode e Thompson (2014). Os conjuntos de dados para teste serão obtidos de três fontes: dados dos exemplos apresentados em Mrode e Thompson (2014), dados simulados e partes de conjuntos de dados reais, obtidos junto ao programa de melhoramento genético de gado de corte GenePlus-Embrapa.

Para realização dos testes apresentados na seção “Resultados e Discussão” foi utilizado um computador de mesa, com 8GB de memoria RAM, com CPU:AMD Phenom(tm) II X4 B95 Processor, utilizando o Ubuntu versão 14.04. Para avaliação do comportamento da implementação baseada em matrizes esparsas, os testes serão realizados em um servidor com maior capacidade de processamento e, em particular, disponibilidade de memória RAM.

Resultados e Discussão

Para implementação da versão da solução de BLUP baseada em matrizes densas, utilizou-se a biblioteca PyPedal tanto para gerar a matriz de parentesco (NRM) quanto sua inversa; enquanto que para montagem do sistema de equações lineares e sua resolução, utilizou-se as bibliotecas SciPy e NumPy (SCIPY, 2015).

Como prova de conceito, essa implementação foi testada utilizando-se o exemplo 3.1 apresentado na referencia (MRODE; THOMPSON, 2014), onde a característica analisada é o ganho de peso à desmama, do inglês Pre-weaning gain – WWG (Tabela 1), assumido-se que a variância genética é dada por $\alpha = 2$. Esses dados são apresentados na Tabela 1.

Tabela 1. Pedigree e ganho de peso à desmama.

Bezerro	Sexo	Pai	Mãe	WWG (kg)
4	Macho	1	Desconhecido	4,5
5	Fêmea	3	2	2,9
6	Fêmea	1	2	3,9
7	Macho	4	5	3,5
8	Macho	3	6	5,0

Fonte: Mrode e Thompson (2014).

O modelo misto utilizado neste exemplo pode ser descrito pelo seguinte conjunto de equações $y_{ij} = p_i + a_j + e_{ij}$, onde y_{ij} é o WWG do j-ésimo bezerro do i-ésimo sexo, p_i é o efeito fixo do i-ésimo sexo, a_j é o efeito aleatório do j-ésimo bezerro e e_{ij} é o erro do efeito aleatório.

O sistema de equações lineares utilizado para estimar os valores dos efeitos estudados é dado por:

$$\begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z + A^{-1}\alpha \end{bmatrix} \begin{bmatrix} b \\ \alpha \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \end{bmatrix}$$

onde A é a matriz de parentesco.

Inicialmente, o programa implementado recebe o Pedigree e a variância genética e, em seguida, monta as matrizes X e Z e, por meio do software PyPedal, obtém a inversa da matriz de parentesco. Por fim, todas essas matrizes são juntadas para formar o sistema de equações lineares acima. Para obter a solução do sistema foi utilizado a implementação do método do gradiente conjugado presente na biblioteca SciPy.

A solução obtida é apresentada na Tabela 2, ilustrando que o resultado coincide com aquele apresentado em Mrode eThompson (2014).

Tabela 2. Solução BLUP (valores estimados dos efeitos genéticos) do problema apresentado na Tabela 1.

Efeitos	Solução
.....Sexo.....	
1 = Macho	4,358
2 = Fêmea	3,404
.....Bezerro.....	
1	-0,041
2	-0,019
3	0,098
4	-0,009
5	-0,186
6	0,177
7	-0,249
8	0,183

Considerações Finais

Até o presente momento, utilizamos Python, suas bibliotecas (SCIPY, 2015) e o pacote PyPedal para construção do sistema linear para solução de BLUP baseado nas matrizes densas geradas por PyPedal (NRM e suas inversas). Essa implementação foi validada por meio de um estudo de caso (dados do exemplo 3.1) em Mrode e Thompson (2014).

No momento, estamos estudando as implementações de PyPedal para geração da NRM e suas inversas para reimplementá-las com base em matrizes esparsas. Em seguida, reimplementaremos a solução do BLUP com base nessas matrizes e avaliaremos a escalabilidade dessa solução em um servidor com grande capacidade de memória RAM.

Referências

BOURDON, R. M. **Understanding animal breeding**. Upper Saddle River: Prentice Hall, 2000. 538 p. ill.

COLE, J. B. **A manual for use of PyPedal**: a software package for pedigree analysis. 2012. Disponível em: <http://pypedal.sourceforge.net/manual/index.html>. Acesso em: 15 out. 2015.

JIANG, J. **Linear and generalized linear mixed models and their applications**. New York: Springer, 2007. 268 p.

MENEZES, G. R. O. **Desenvolvimento e implementação de metodologias genético-estatísticas em avaliações genéticas de gado de corte**. Campo Grande: Embrapa Gado de Corte, 2015. (Embrapa. Macroprograma 2). Código SEG: 02.13.14.004.00.00. Projeto em andamento.

MISZTAL, I. **Computational techniques in animal breeding**. Athens: University of Georgia, 2011. 188 p. ill. Disponível em: <http://nce.ads.uga.edu/~ignacy/course11.pdf>. Acesso em: 15 out. 2015.

MRODE, R. A.; THOMPSON, R. **Linear models for the prediction of animal breeding values**. 3rd ed. Boston: CABI, 2014. 360 p.

PYTHON SOFTWARE FOUNDATION. **Python**. 2015. Disponível em: <https://www.python.org/>. Acesso em: 15 out. 2015.

ROSA, A. do N.; MARTINS. E. N.; MENEZES, G. R. de O.; SILVA, L. O. C. da (Ed.). **Melhoramento genético aplicado em gado de corte**: Programa Geneplus-Embrapa. Brasília, DF: Embrapa; Campo Grande, MS: Embrapa Gado de Corte, 2013. 241 p.

SCIPY. **Numpy and Scipy documentation**. Disponível em: <http://docs.scipy.org/doc/>. Acesso em: 15 out. 2015.

WATKINS, D. S. **Fundamentals of matrix computations**. 3rd ed. Hoboken: Wiley, 2010. 644 p.