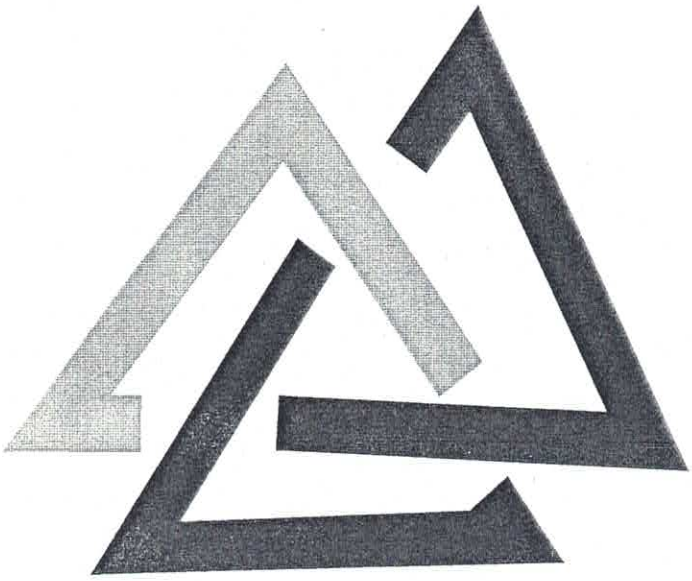
#120

Support vector machine e support vector regression para seleção de SNPs em GWAS

Fabrízzio Condé de Oliveira, Fernanda Nascimento Almeida, Priscila V. S. Z. Capriles, Wagner Arbex, Carlos Cristiano H. Borges

Resumo: Métodos comumente usados para estudos de associação em escala genômica ou genome-wide association studies (GWAS) avaliam separadamente o impacto de cada marcador do tipo SNP (single nucleotide polymophism) no fenótipo considerado. Essa avaliação pode ser feita por meio de regressão linear simples para o caso de fenótipo contínuo, ou por meio de regressão logística para fenótipo dicotômico mapeado por caso-controle. Um possível cenário é quando o fenótipo é determinado pela influência de vários genes com pequenos efeitos aditivos sobre o mesmo. Outra situação é quando dois ou mais SNPs pouco significativos isoladamente demonstram grande efeito de interação, gerando padrões complexos na relação genótipo-fenótipo (epistasia). Atualmente, os chips de alta densidade possuem aproximadamente 800.000 marcadores genéticos e o número de indivíduos nas amostras consideradas em GWAS estão na ordem de centenas ou poucos milhares, gerando uma limitação para aplicação de diversas técnicas multivariadas. Desta forma, para possibilitar a seleção simultânea de vários SNPs, onde há o desequilíbrio entre o tamanho da amostra e a quantidade de marcadores no conjunto de dados, pode-se usar as técnicas de Máquina de Vetor Suporte (SVM - Support Vector Machine) para problemas de classificação e para problemas de regressão (SVR - Support Vector Regression). O presente trabalho tem por objetivo propor um método baseado em um modelo multiatributo, construído a partir de filtros estatísticos tradicionais em GWAS juntamente com SVM ou SVR baseado nos kernels linear, radial e Pearson Universal. O conjunto de dados real é composto de 240 touros da raça GIR cujo genótipo possui 56.947 marcadores SNPs genotipados por chips da Illumina 50kv2. O fenótipo é composto pelo PTA (Predicted Transmitting Ability) do leite de cada animal, sendo o mesmo uma variável contínua. Para avaliar o SVM, discretizou-se o PTA do leite em duas classes: uma abaixo ou igual a mediana (rótulo 0) e outra acima da mediana (rótulo 1). Os filtros de controle de qualidade para os SNPs aplicados foram call rate >= 95%, minor allele frequency (MAF) >= 5% e equilíbrio de Hardy-Weinberg >= 0,05/56.947 (5% com correção de Bonferroni). Após o filtro restaram 22.844 marcadores. O próximo passo foi usar o p-valor bruto do coeficiente de correlação de Spearman entre o marcador e o PTA para ajustá-los por meio da correção de Bonferroni. A partir do cálculo do p-valor ajustado, o mesmo foi usado para ordenar os SNPs em 17 grupos os quais formaram uma sequência crescente de subconjuntos de SNPs. A medida de acurácia para os modelos com SVM (classificação) foi a área abaixo da curva ROC (Receiver Operating Characteristic), enquanto que para os modelos SVR (regressão), adotou-se o coeficiente de correlação de Pearson. Foi realizada uma validação cruzada com 10 folds, o que permite obter apenas uma medida de acurácia. Repetiu-se esse processo 10 vezes, no intuito de generalizar os resultados pela média e avaliar a precisão através do desvio padrão objetivando a comparação dos modelos. O método de seleção de marcadores SNPs com base no SVR demonstrou bom desempenho em comparação com métodos comumente usados em GWAS. Isso ocorreu porque a maior correlação média do SVR com kernel radial foi 0,80 para o grupo de marcadores que possui 409 marcadores com p-valor menor que 0,80, enquanto que para o grupo de p-valor menor que 10e-7, que contem somente 3 marcadores, a correlação média foi 0,57.



SIMMEC EMMCOMP 2014

XI Simpósio de Mecânica Computacional
II Encontro Mineiro de Modelagem Computacional
Juiz de Fora, 28-30 de Maio de 2014, UFJF