Software for detecting gene-gene interactions using Multifactor Dimensionality Reduction: Introducing the R package MDR

Stacey J. Winham[1] and Alison A. Motsinger-Reif[1,2]

[1]Department of Statistics, [2]Bioinformatics Research Center

North Carolina State University, Raleigh NC 27695

October 15, 2010

North Carolina State University Department of Statistics Technical Reports #2632

*Address for Correspondence:

Stacey J. Winham
Department of Statistics,
North Carolina State University,
Raleigh NC, 27695-7566, USA
EMAIL: sjwood@statncsu.edu

Alison A Motsinger-Reif, Ph.D.
Bioinformatics Research Center,
Department of Statistics,
1 Lampe Dr, CB 7566
North Carolina State University,
Raleigh NC, 27695-7566, USA
TEL: 919-515-3574
FAX: 919-515-7315
EMAIL: motsinger@stat.ncsu.edu

Introduction

With advances in genotyping technologies, a breadth of high-dimensional data is now available with unprecedented numbers of genetic markers to perform association mapping in human genetics. Identifying variants associated with complex human traits is a common problem and data-mining approaches to variable selection have become commonplace methods of analysis. There is growing evidence that epistasis may play a role in disease risk, and many variable selection approaches have been developed within the bioinformatics community to consider potential gene-gene and gene-environment interactions. One of the most commonly used techniques to identify interactions for case-control data is Multifactor Dimensionality Reduction (MDR). MDR is a nonparametric exhaustive search method that considers all combinations of potentially interacting loci [Ritchie, et al. 2001]. MDR traditionally creates a classification rule using a Naïve Bayes classifier, assigning genotype combinations with a large ratio of cases to controls as high-risk and low-risk otherwise. Using this high-risk/low-risk parameterization, a measure of the accuracy of the classification rule is evaluated and a final model is chosen to maximize this accuracy.

Currently software is available which implements the MDR method, including a GUI implementation available at http://www.epistasis.org [Hahn, et al. 2003]. We introduce a package for the R statistical language to implement the method. This package is designed to provide an alternative implementation for R users; the package has great flexibility and utility for both data analysis and research. This package implements the MDR method for variable selection of

interactions as first outlined in [Ritchie, et al. 2001], providing options for internal validation and functions to summarize the fit and perform post-hoc inference. This package is available through the CRAN website (http://cran.r-project.org/) or through the following website: www4.stat.ncsu.edu/~motsinger.

Implementation

This package utilizes balanced accuracy, the arithmetic mean of sensitivity and specificity, as the evaluation measure for comparing different combinations of variables [Velez, et al. 2007]. Other evaluation measures are possible, including additional contingency table measures such as normalized mutual information (NMI), as noted in [Bush, et al. 2008] but are not included in this package. This package assumes binary case-control data with categorical predictor variables. The binary response variable is coded as 0 or 1, and the categorical predictors (typically SNP genotypes) are coded numerically (0, 1, 2, etc.). The user can specify the particular genotype encoding. Additionally, the threshold for assigning high-risk/low-risk status to variable combinations can also be controlled by the user.

Internal Validation

In all data-mining methods, over-fitting a model to the particular data set at hand is a concern. It is suggested that the MDR method is implemented in conjunction with an internal validation technique and this package provides two such procedures to fit an MDR model and select loci: *k*-fold cross-validation and three-way split internal validation.

In *k*-fold cross-validation, the data is randomly split into *k* equal intervals, with where *k*-1 intervals are used for training and one interval is used for testing [Motsinger and Ritchie 2006]. The best MDR model is determined from the training set for each size of interaction and an estimate of the model's prediction accuracy is calculated from the testing set. This procedure is repeated for all *k* possible splits of the data and a final model is chosen to maximize both prediction accuracy and cross-validation consistency across each split. The function 'mdr.cv' implements cross-validation and allows the user to specify the highest level of interaction to consider, as well as the number of intervals *k*; typically a value of *k*=5 or 10 yields high performance [Motsinger and Ritchie 2006].

In three-way split internal validation, the data is randomly split into 3 sets for training, testing, and validation [Winham, et al. 2010]. MDR is first implemented in the training set for all possible combinations of loci. The *x* models with the highest balanced accuracy in the training set are retained for evaluation in the testing set. MDR is next performed on all *x* models in the testing set and the best model for each level of interaction is preserved for evaluation of predictive ability in the validation set. A final model is chosen to maximize the balanced accuracy in the validation set. The function 'mdr.3WS' implements three-way split internal validation and allows the user to specify the ratio of the three data splits (training:testing:validation), and also the number of potential models *x* from the training set to be evaluated in the testing set.

Both internal validation methods create objects of class 'mdr', which is a list of the final selected model loci and its prediction accuracy, the top models and their prediction accuracies, and the high-risk/low-risk characterization of the final model.

Methods

Three methods exist for objects of class 'mdr': 'summary', 'plot', and 'predict'. The 'summary' method provides a table summarizing the model fit at each stage of interaction. The 'plot' method provides a contingency table of bar graphs for the final model, portraying the numbers of cases and controls in each genotype combination, similar to the GUI implementation at http://www.epistasis.org. The 'predict' method allows the user to predict case-control status on a new, independent set of data with a model obtained from a previously fit 'mdr' object.

Post-hoc functions for inference

After an MDR model has been fit, a number of functions exist for inference on that fit. Permutation testing is available to test the significance of the reported measure of prediction accuracy; case-control status is randomly permuted a number of times (specified by the user), and the resulting prediction accuracies from each MDR fit of the permuted data sets are compared to a specified accuracy [Motsinger-Reif 2008]. Additionally, estimates of prediction accuracy are obtained from retrospective case-control data, and therefore may not reflect the true accuracy of prospective predictions. Using a previously estimated population prevalence rate provided by the user, these prediction accuracy estimates can be adjusted using one of two

available post-hoc procedures implemented in 'boot.error' and 'mdr.ca.adj'

[Winham and Motsinger-Reif 2010].

References

Bush WS, Edwards TL, Dudek SM, McKinney BA, Ritchie MD. 2008. Alternative contingency table measures improve the power and detection of multifactor dimensionality reduction. Bmc Bioinformatics 9.

Hahn LW, Ritchie MD, Moore JH. 2003. Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions. Bioinformatics 19(3):376-382.

Motsinger-Reif AA. 2008. The effect of alternative permutation testing strategies on the performance of multifactor dimensionality reduction. BMC Res Notes 1:139.

Motsinger AA, Ritchie MD. 2006. The effect of reduction in cross-validation intervals on the performance of multifactor dimensionality reduction. Genet Epidemiol 30(6):546-55.

Ritchie MD, Hahn LW, Roodi N, Bailey LR, Dupont WD, Parl FF, Moore JH. 2001. Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. Am J Hum Genet 69(1):138-47.

Velez DR, White BC, Motsinger AA, Bush WS, Ritchie MD, Williams SM, Moore JH. 2007. A balanced accuracy function for epistasis modeling in imbalanced datasets using multifactor dimensionality reduction. Genet Epidemiol 31(4):306-15.

Winham SJ, Motsinger-Reif AA. 2010. The Effect of Retrospective Sampling on Estimates of Prediction Error for Multifactor Dimensionality Reduction. Ann Hum Genet.

Winham SJ, Slater AJ, Motsinger-Reif AA. 2010. A comparison of internal validation techniques for multifactor dimensionality reduction. Bmc Bioinformatics 11:394.