Genetic Epidemiology 33 (Supplement 1): S51–S57 (2009)

Machine Learning in Genome-Wide Association Studies

Silke Szymczak,[1]* Joanna M. Biernacka,[2] Heather J. Cordell,[3] Oscar González-Recio,[4] Inke R. König,[1] Heping Zhang,[5] and Yan V. Sun[6]

[1]*Institut für Medizinische Biometrie und Statistik, Universität zu Lübeck, Lübeck, Germany*
[2]*Department of Health Sciences Research, Mayo Clinic, Rochester, Minnesota*
[3]*Institute of Human Genetics, International Centre for Life, Newcastle University, Central Parkway, Newcastle upon Tyne, United Kingdom*
[4]*Department of Dairy Science, University of Wisconsin-Madison, Madison, Wisconsin*
[5]*Department of Epidemiology and Public Health, Yale University School of Medicine, New Haven, Connecticut*
[6]*Department of Epidemiology, School of Public Health, University of Michigan, Ann Arbor, Michigan*

Recently, genome-wide association studies have substantially expanded our knowledge about genetic variants that influence the susceptibility to complex diseases. Although standard statistical tests for each single-nucleotide polymorphism (SNP) separately are able to capture main genetic effects, different approaches are necessary to identify SNPs that influence disease risk jointly or in complex interactions. Experimental and simulated genome-wide SNP data provided by the Genetic Analysis Workshop 16 afforded an opportunity to analyze the applicability and benefit of several machine learning methods. Penalized regression, ensemble methods, and network analyses resulted in several new findings while known and simulated genetic risk variants were also identified. In conclusion, machine learning approaches are promising complements to standard single-and multi-SNP analysis methods for understanding the overall genetic architecture of complex human diseases. However, because they are not optimized for genome-wide SNP data, improved implementations and new variable selection procedures are required. *Genet. Epidemiol.* 33 (Suppl. 1):S51–S57, 2009.

Key words: Genetic Analysis Workshop; data mining; penalized regression; random forests; network analysis

Contract grant sponsor: The German Ministry of Education and Science; Contract grant number: 01GS0831; Contract grant sponsor: NIH; Contract grant number: R01 GM031575.
*Correspondence to: Silke Szymczak, Institut für Medizinische Biometrie und Statistik, Universität zu Lübeck, Maria-Goeppert-Str. 1, 23562 Lübeck, Germany. E-mail: silke.szymczak@imbs.uni-luebeck.de
Published online in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/gepi.20473

INTRODUCTION

In recent years, genome-wide association studies (GWAS) have been proven to be successful in the identification of new genetic variants that influence the risk of a wide range of complex diseases, including cardiovascular [Mohlke et al., 2008; Samani et al., 2007] and autoimmune diseases [Lettre and Rioux, 2008] as well as cancer [Easton and Eeles, 2008]. Most GWAS focus on the detection of main effects by using an allele- or genotype-based test for each single-nucleotide polymorphism (SNP) separately. However, the identified genetic effects tend to be moderate and explain only a small fraction of the overall heritability [Frazer et al., 2009]. Because multiple interacting genetic loci in combination with environmental risk factors are expected to contribute to susceptibility to disease, multiple SNPs and interaction effects should be analyzed simultaneously.

However, there are several challenges in studying the joint effects of multiple genetic and environmental variables. First, in typical GWAS, genotypes of up to one million SNPs are determined in several thousand subjects, leading to the small n, large p problem (many more variables (SNPs) than samples). Second, when a large number of SNPs are genotyped on a genome-wide scale, linkage disequilibrium (LD) between SNPs (resulting in correlated variables) needs to be taken into account. For these reasons, standard multi-variable statistical approaches like multiple linear or logistic regression are not well suited for genome-wide data. Machine learning algorithms provide several alternatives for performing multi-SNP analyses. For instance, penalized regression methods extend standard regression techniques so that a large number of possibly correlated variables may be analyzed. Relevant SNPs are identified based on regression coefficient estimates, and interactions may be modeled explicitly. In contrast, nonparametric approaches like ensemble methods or neural networks can be used to model complex relationships between variables without the need to specify a particular model. Some of these methods provide variable ranking methods to select the most important SNPs for predicting the outcome variable.

The Genetic Analysis Workshop (GAW) 16 provided three data sets that were used to analyze the applicability of several machine learning methods for GWA data and to identify context-specific solutions for method-inherent problems (see Table I for an overview of group contributions). Genome-wide SNP data were made available as the first data set by the North American Rheumatoid Arthritis Consortium (NARAC), a case-control study of rheumatoid arthritis (RA) [Amos et al., 2009]. The second GWA data set and a series of cardiovascular- and diabetes-related phenotypes were provided by Framingham Heart Study

TABLE I. Overview of group contributions (ordered by first author)

First author	Data set	Phenotype	Method	Question
Arshadi	NARAC	RA	GBM	Prediction
Croiseau	NARAC	RA	Group LASSO	Main effects
D'Angelo	NARAC	RA	LASSO	Gene-gene interactions
González-Recio	NARAC	RA	Bayesian LASSO	Gene-gene interactions
Kim	FHS (simulated)	MI, CAC	RF	Main effects
Schwarz	FHS	–	RF	Genotype imputation
Stassen	NARAC	IgM	ANN	Main effects
Sun	NARAC	Anti-CCP	RR	Main effects
Tang	NARAC	RA	RF	Main effects, haplotypes, gene-gene interactions
Wang	NARAC	RA	RF	Main effects
Woo	NARAC	RA	ANN, SVM, kNN	Prediction
Yang	FHS (simulated)	CAC	RF	Main effects
		MI, CAC	BNT	Causal relationship

NARAC, North American Rheumatoid Arthritis Consortium; FHS, Framingham Heart Study; RA, rheumatoid arthritis (dichotomous); MI, myocardial infarction (dichotomous); CAC, coronary artery calcification (continuous); IgM, immunoglobulin M (categorical); Anti-CCP, anti-cyclic citrullinated peptide (continuous); GBM, gradient boosting machine; LASSO, least absolute shrinkage and selection operator; RF, random forest; ANN, artificial neural network; RR, ridge regression; SVM, support vector machine; kNN, *k*-nearest neighbor; BNT, Bayesian network analysis.

(FHS), a community-based longitudinal study of three generations [Cupples et al., 2009]. In the third data set, several gene-gene and gene-environment interactions were simulated based on genotypic data from FHS [Kraja et al., 2009].

Because ensemble methods, and especially random forests (RF), were discussed in detail in a GAW15 summary paper [Ziegler et al., 2007] on data mining techniques, this contribution focuses mainly on penalized regression methods. New developments and applications of ensemble and network methods are then briefly summarized. The remainder of this article is organized as follows: After an introductory section about the utilized methods, we present the main results of our group's contributions followed by a discussion of method-specific problems and possible solutions.

PENALIZED REGRESSION

In classical linear regression, the quantitative response variable Y is modeled as a linear combination of the predictor variables $X_1, \ldots, X_p$:

$$Y = \beta_0 + \sum_{j=1}^{p} \beta_j X_i,$$

where β_0 and $\beta = (\beta_1, \ldots, \beta_p)^T$ denote intercept and regression coefficients. This model is fitted using a training data set consisting of the observations $(x_1, y_1), \ldots, (x_n, y_n)$ of n samples. Each $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^T$ denotes a vector of p observations for the predictor variables of sample i. The values of the unknown parameter β_i may be estimated by minimizing the residual sum of squares

$$\begin{aligned}\hat{\beta} &= \arg\min_{\beta} g(\beta_0, \beta) \\ &= \arg\min_{\beta} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2.\end{aligned}$$

The predictor variables may be SNPs (e.g., coded as 0, 1, or 2 according to the number of risk alleles), environmental risk factors, or a combination of both. Gene-gene or gene-environment interactions may be modeled by adding interaction terms to the regression equation.

Because classical regression approaches require the number of samples to exceed the number of variables, they are not applicable in case of GWA data. Additionally, least-squares estimates of regression coefficients may be highly unstable, especially in cases of correlated predictor variables, which lead to low prediction accuracy.

To overcome these challenges, penalized regression approaches, also called shrinkage methods, were proposed. Although shrinking (e.g., setting some of the regression coefficients to zero) may result in biased estimates, these estimates often have a smaller variance. As a consequence, the prediction accuracy is improved due to a smaller mean square error [Hastie et al., 2001]. Additionally, these approaches facilitate variable selection because only important predictor variables remain in the model. Regression coefficients are shrunk by imposing a penalty on their size. Specifically, these methods minimize an expanded function

$$f(\beta_0, \beta) = g(\beta_0, \beta) + h(\delta, \beta),$$

with $h(\delta, \beta)$ denoting a penalty function with the tuning parameter δ.

For application in case-control data with a binary outcome variable, penalized regression approaches can be modified by replacing the residual sum of squares by the negative log-likelihood function in the logistic regression framework. In the following, approaches using different penalty functions will be presented focusing on methods that were applied by contributors in our group.

RIDGE REGRESSION

In ridge regression [Hoerl and Kennard, 2000], the size of the coefficients is constrained by the L_2 penalty $\sum_{j=1}^{p} \beta_j^2 \leq s$. Equivalently, the corresponding estimates

minimize the penalized residual sum of squares

$$\hat{\beta}^{\text{ridge}} = \arg\min_{\beta} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \delta \sum_{j=1}^{p} \beta_j^2.$$

The tuning parameter δ controls the amount of shrinkage. If it is set to zero, the estimated ridge regression coefficients are equivalent to the classical coefficient estimates. Otherwise, a larger value of δ corresponds to a larger amount of shrinkage. This minimization problem can be solved analytically. In an application of differentiating the potential causal from noncausal SNPs, ridge regression showed an advantage over the regular multiple regression method and single-locus analysis in genetic regions with strongly correlated SNPs [Malo et al., 2008].

LASSO

In contrast to ridge regression, least absolute shrinkage and selection operator (LASSO) [Tibshirani, 1996] uses the L_1 penalty $\sum_{j=1}^{p} |\beta_j| \leq t$.

The resulting regression problem

$$\hat{\beta}^{\text{LASSO}} = \arg\min_{\beta} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \delta \sum_{j=1}^{p} |\beta_j|$$

is nonlinear in the y_i, and a quadratic programming algorithm is used to estimate the regression coefficients. If δ is small, some of the coefficients will be exactly zero so that only relevant variables remain in the model. Wu et al. [2009] applied LASSO to genome-wide SNP data for both marginal and interaction predictors. Features of tuning parameter selection, predictor selection, and false-discovery rate for global significance were incorporated within a fast computing implementation.

GROUP LASSO

For the group LASSO [Yuan and Lin, 2006], predictor variables are divided into G groups. A group-wise penalty is applied, leading to the following regression problem:

$$\hat{\beta}^{\text{group LASSO}} = \arg\min_{\beta} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \delta \sum_{g=1}^{G} \sqrt{\sum_{j \in g} \beta_j^2},$$

which is solved with an iterative algorithm. This penalty is an intermediate between the penalties used in ridge regression and LASSO. Moreover, group LASSO corresponds to LASSO if each group consists of a single variable. Group LASSO has the advantage that variables in a group can only be selected together. This property can be used in genetic association studies, for instance, to model interactions only when main effects are present or to group SNPs in the same gene or pathway.

MAIN RESULTS

Several penalized regression approaches were applied to identify genetic variants and gene-gene interactions that are associated with RA or an intermediate phenotype. Sun et al. [2009] used ridge regression to detect SNPs associated with variations of anti-cyclic citrullinated peptide levels, a clinical predictor of RA development. Incorporating information about multiple correlated genetic variants led to the identification of an SNP near the *HLA-B* gene that was not significant in single-SNP analyses. In contrast, Croiseau and Cordell [2009] observed no advantage of group LASSO over the standard trend test for the detection and localization of SNPs associated with RA.

However, D'Angelo et al. [2009] found that the identification of gene-gene interactions was more successful. By combining LASSO in a logistic regression framework with principal-component analysis for dimension reduction, they identified two significant gene-gene interactions within the major histocompatibility complex on chromosome 6p that were also found using alternative approaches. González-Recio et al. [2009] searched for gene-gene interactions among SNPs in the HLA region and between an HLA SNP and an SNP elsewhere in the genome using a Bayesian threshold LASSO [Park and Casella, 2008]. Many of the SNPs with identified main and interaction effects were in genes known to be associated with RA.

SPECIFIC PROBLEMS AND SOLUTIONS

It was computationally infeasible to apply penalized regression methods on genome-wide data. For this reason, sets of possibly interesting SNPs were selected based on previous findings or by statistical methods. Sun et al. [2009] and Croiseau and Cordell [2009] restricted their analyses to SNPs in regions on chromosomes 1, 6, and 9 with known susceptibility loci for RA [Plenge et al., 2007]. Similarly, D'Angelo et al. [2009] focused on SNPs in RA candidate genes on chromosome 6. In contrast, González-Recio et al. [2009] applied an information-gain criterion and a wrapper procedure to select SNPs for further analysis.

Instead of P-values for each SNP, penalized regression approaches give an estimated regression coefficient for each variable. To select relevant SNPs, Croiseau and Cordell [2009] estimated standard errors and corresponding confidence intervals using a bootstrap method with a normality assumption. The covariance matrix can be calculated numerically if ridge regression is applied. Additionally, Sun et al. [2009] corrected for the bias of the estimated coefficients by restandardizing the Z-scores using the slope in a quantile-quantile plot of the observed Z-scores and the standard normal quantiles.

The variable selection procedure strongly depends on the tuning parameter λ for the amount of shrinkage. This parameter is often chosen based on computationally intensive cross-validation techniques. However, less time-consuming approaches were preferred for the analysis of SNP data. Sun et al. [2009] used several possible values and ranked the SNPs based on the maximal Z-score of all considered models. Croiseau and Cordell [2009] set λ to the logarithm of the group size as proposed by Meier et al. [2008]. The Bayesian approach used by González-Recio et al. [2009] considers posterior distributions to provide probabilities of coefficient estimates being different from zero. Furthermore, it allows setting the parameter λ as unknown in the sampling process.

ENSEMBLE METHODS

Ensemble methods use a set of classifiers or regression functions. Predictions of these so-called base learners are

combined by weighted voting to generate the overall prediction of the ensemble. It has been shown that these ensembles perform better than their individual components under certain conditions [Dietterich, 2000]. First, the components have to be weak learners, i.e., they are better than random guessing. Often, nonparametric classification and regression trees (CART) [Breiman et al., 1984] are used as base learners. These trees do not require the specification of a particular linear or nonlinear relationship between predictor and response variables. Second, the predictions of the base learners have to be different. In principle, different learning algorithms, such as decision trees, support vector machines, and discriminant analysis, may be used to construct distinct base learners. However, we will focus on approaches that use base learners of the same type that are trained on slightly different data sets. One popular and very general method is bagging (short for bootstrap aggregating) [Breiman, 1996] that draws bootstrap samples out of the original data. A special version is RF [Breiman, 2001], for which a CART is grown on each bootstrap sample. RF includes an additional random component in the learning process, namely, that the variables are randomly selected to determine the optimal split at each node of the tree. RF has been utilized in various studies to predict disease status using SNPs [Sun et al., 2007], to rank SNP predictors [Schwarz et al., 2007; Sun et al., 2008], and to identify the epistatic effects related to human diseases [García-Magariños et al., 2009].

Another method to generate an ensemble is boosting, which was first used in a version of the well known AdaBoost algorithm [Freund and Schapire, 1997]. Here, each base learner is constructed using a reweighted data set, with the weights depending on the results of the previous base learner. Boosting can also be interpreted as the steepest descent algorithm in a function space [Friedman et al., 2000]. The gradient boosting machine (GBM) [Friedman, 2001] minimizes a context-dependent loss function using gradient descent and trees as base learners. Recently, Wan et al. [2009] used boosting of CARTs combined with a hierarchical learning approach to identify multi-SNP interactions in GWA data.

The RF and GBM approaches provide variable importance measures that can be used to select the most relevant predictors. As an advantage over *P*-values of single-SNP tests, these measures implicitly incorporate interaction effects. Two different methods are commonly used to define importance measures in RF. The first one determines the trees' improvement in the Gini splitting criterion for each variable and is therefore denoted as Gini importance (GI) in this paper. The permutation importance (PI) is based on the difference in prediction accuracy before and after permuting all values of the variable so that any association with the response variable is destroyed.

MAIN RESULTS

Kim et al. [2009] and Yang and Gu [2009] applied RF to the simulated data set. Kim et al. [2009] were able to identify the environmental risk factors for the binary phenotype myocardial infarction (MI) as well as for the quantitative variable coronary artery calcification (CAC) but only one SNP that interacts with smoking for the risk of MI seemed to be relevant in their RF analysis. Similar results were observed by Yang and Gu [2009], who mainly detected the environmental risk factors and only one SNP that influences CAC.

In contrast, Tang et al. [2009] and Wang et al. [2009] identified many known and several new SNPs that appear to contribute to RA risk. Using a new permutation approach, Tang et al. [2009] also searched for gene-gene interactions but they were not able to find strong evidence for these kinds of effects.

Schwarz et al. [2009] used genotype data of FHS in combination with HapMap CEU samples to evaluate the internal method of RF for imputing missing genotypes. They concluded that alternative approaches like IMPUTE [Marchini et al., 2007] are more accurate for imputing untyped SNPs.

In contrast to the aforementioned applications of RF, Arshadi et al. [2009] used a GBM to analyze the effect of population stratification on prediction accuracy. Their proposed approach clustered probands on the axes of genetic variation and subsequently developed a GBM for each cluster separately. The resulting model had a higher prediction performance in comparison with a model confounded by ethnicity.

SPECIFIC PROBLEMS AND SOLUTIONS

Several aspects of variable importance measures were analyzed by contributors to Group 8b. Kim et al. [2009] showed that causal SNPs and important covariates were more frequently included in the list of top-ranking variables when GI was used instead of PI for evaluating the variable importance. Tang et al. [2009] proposed new gene-level importance measures based on scaled PI. A gene importance measure was defined as the maximum or mean importance of all SNPs in the corresponding gene. Haplotype importance measures were calculated in a similar way based on forests that were built on random samples of haplotypes. To identify interactions between two genes *a* and *b*, they examined changes in variable importance for gene *a* when genotypes of SNPs in gene *b* were permuted and vice versa. Arshadi et al. [2009] used the relative influence measure of GBM to compare high ranked SNPs in their GBMs that incorporate ethnicity in different ways. Similar to the importance scores of RF, SNPs with large influence scores do not necessarily have to be selected based on *P*-values of single-SNP analyses.

One major challenge of the importance measures is that calculation of *P*-values is not straightforward. For example, the significance test for PI as proposed by Breiman and Cutler [2009] has several undesired statistical properties [Strobl and Zeileis, 2008]. In contrast, Wang et al. [2009] generated permuted data with no association between predictor variables and RA and recalculated the variable importance on the permuted data, resulting in empirical *P*-values. To avoid fitting to noisy predictor variables, Yang and Gu [2009] used an iterative variable selection procedure, resulting in an RF that is built on a small set of highly important variables. This approach selected the true causal SNPs and covariates more often than a standard RF analysis using all predictor variables.

RF analyses using the default value of *mtry* for classification performed better in the selection of true predictor variables [Kim et al., 2009] than very small *mtry* values. In contrast, Díaz-Uriarte and Alvarez de Andrés [2006] observed that different values of *mtry* led to similar

results. However, in situations with very few relevant predictor variables among many potential predictors, they observed that small *mtry* values will select only uninformative variables for incorporation in many trees, resulting in an increased error rate.

Arshadi et al. [2009] showed that the GBM model is not strongly influenced by the number of irrelevant SNPs included in the model. In contrast, Yang and Gu [2009] observed that true risk SNPs were less often in the top-ranked variables when the number of noise SNPs was increased.

NETWORKS

The concept of a network is used in completely different ways by the two approaches that will be described in this subsection. On the one hand, an artificial neural network (ANN) is a black box method that can be used for classification and regression problems (see [Ripley, 1996] for an introduction). On the other hand, Bayesian network analysis (BNT) is used to model complex relationships between variables from empirical data.

ARTIFICIAL NEURAL NETWORKS

ANN can be used to model complex nonlinear relationships between predictor and response variables. These methods are well suited for problems with a large signal-to-noise ratio when the primary objective is prediction rather than selection of relevant predictor variables. They try to emulate the biological network of neurons in the brain by connecting the predictor and the response variables using layers of intermediate (hidden) nodes, possibly with feedback connections. In each of these nodes a weighted sum of the corresponding input nodes is calculated and after a transformation, forwarded to the nodes in the next layer. Training of an ANN involves modifying the weights according to a learning algorithm using a training data set. Feed-forward neural networks have a rather simple topology with one or more layers of hidden nodes but without any feedback loops. A popular learning strategy for this kind of network is the back-propagation algorithm [Rumelhart et al., 1986] that uses the method of gradient descent to determine optimal weights. An ANN was used in a candidate gene study to identify SNPs with major effects as well as two- and three-way interactions [Tomita et al., 2004].

BAYESIAN NETWORK ANALYSIS

A BNT describes the joint distribution of predictor and response variables graphically, with nodes representing the variables and edges denoting dependencies and conditional independencies [Pearl, 1988]. This representation captures multiple associations and interactions between SNPs and the phenotype as well as between SNPs due to LD simultaneously. A Bayesian approach is applied to select the most probable network given the empirical data (for a tutorial on BNT see Heckerman [1999]). Directions of edges in the resulting network are supported by the data but they do not need to indicate causality. In a candidate gene study, complex gene-gene interactions were identified by a BNT that was subsequently used for prognosis [Sebastiani et al., 2005].

MAIN RESULTS

Stassen et al. [2009] analyzed the reproducibility of ANN predictors across populations and across SNP sets. An ANN classifier using 15 genomic loci was built with GAW15 data [Amos et al., 2007] as a training sample. Results of a similar analysis of the GAW16 RA data as a test set in combination with a "competitive SNP set" approach overlapped only in part with these loci. Woo et al. [unpublished] used ANN in addition to two other popular supervised learning algorithms, support vector machines [Vapnik, 1996] and *k*-nearest neighbor [Dasarathy, 1991], to study the effect of differential SNP encoding on the classification accuracy. Coding each SNP according to its mode of inheritance resulted in increased prediction performance independent of the selected algorithm.

Yang and Gu [2009] used the simulated data set to analyze the ability of BNT to detect relationships between known predictor variables and the binary MI event and its intermediate phenotype CAC. Only some of the true relationships could be recovered in the BNT analysis, especially if a large number of noise SNPs were also included in the model process.

SPECIFIC PROBLEMS AND SOLUTIONS

Both Stassen et al. [2009] and Woo et al. [unpublished] selected a simple feed-forward neural network with a single hidden layer consisting of a small number of nodes and a single output variable. However, the number of nodes of the hidden layer varied relative to the number of predictors. Stassen et al. [2009] used the same number, whereas Woo et al. [unpublished] reduced the number of nodes of the hidden layer to about 1/20 of the predictors. In addition, to get an approximately unbiased estimate of the prediction error, 10-fold cross validation was used by both groups

DISCUSSION AND CONCLUSION

Table I shows that contributions of our GAW16 Group 8b applied a wide range of different machine learning methods to model the effect of many SNPs on the susceptibility to complex diseases simultaneously. Several qualitative and quantitative phenotypes were used to detect main effects and gene-gene interactions. Additionally, contributors focused on evaluations of prediction accuracies, genotype imputation, and models of causal relationships.

Penalized regression and ensemble methods as well as network analysis were able to detect many known genetic risk variants for RA. In addition, several new SNPs associated with RA were identified that were not detected with standard single-SNP analyses. Finding true relationships in the simulated data set turned out to be difficult because many main effects were very small so that even standard trend tests or simple linear regression do not result in genome-wide significance [Kim et al., 2009].

However, existing implementations of machine learning methods pose several limitations for application to genome-wide data. Penalized regression methods are not able to deal with all SNPs of a GWA simultaneously, underlining the need for improved implementations. In contrast, RFs and other ensemble methods provide variable importance measures that can be directly applied to genome-wide data and combined with standard single-SNP tests for screening purposes. However, further research is needed to identify

and evaluate variable selection procedures that are especially suited for genetic data.

In conclusion, despite current limitations, machine learning methods are sensible complements to standard single-SNP tests for unraveling the genetic basis of complex diseases.

ACKNOWLEDGMENTS

The authors thank all GAW16 Group 8b participants for interesting discussions and helpful comments. The work of Silke Szymczak and Inke R. König was supported by the German Ministry of Education and Science, grant 01GS0831, and an intramural grant of the University at Lübeck, Germany. The Genetic Analysis Workshops are supported by NIH grant R01 GM031575 from the National Institute of General Medical Sciences.

REFERENCES

Amos CI, Chen WV, Remmers E, Siminovitch KA, Seldin MF, Criswell LA, Lee AT, John S, Shephard ND, Worthington J, Cornelis F, Plenge RM, Begovich AB, Dyer TD, Kastner DL, Gregersen PK. 2007. Data for Genetic Analysis Workshop (GAW) 15 Problem 2, genetic causes of rheumatoid arthritis and associated traits. BMC Proc 1:S3.

Amos CI, Chen WV, Seldin MF, Remmers E, Taylor KE, Criswell LA, Lee AT, Plenge RM, Kastner DL, Gregersen PK. 2009. Data for Genetic Analysis Workshop 16 Problem 1, association analysis of rheumatoid arthritis data. BMC Proc 3:S2.

Arshadi N, Chang B, Kustra R. 2009. Predictive modeling in case-control single-nucleotide polymorphism studies in the presence of population stratification: a case study using Genetic Analysis Workshop 16 Problem 1 dataset. BMC Proc 3:S60.

Breiman L. 1996. Bagging predictors. Mach Learn 24:123–140.

Breiman L. 2001. Random forests. Mach Learn 45:5–32.

Breiman L, Cutler A. 2009. Random forests—Classification manual. URL: http://www.math.usu.edu/~adele/forests/, last access 10/20/2009.

Breiman L, Friedman J, Olshen R, Stone C. 1984. Classification and Regression Trees. Belmont, CA: Wadsworth.

Croiseau P, Cordell HJ. 2009. Analysis of North American Rheumatoid Arthritis Consortium data using a penalized logistic regression approach. BMC Proc 3:S61.

Cupples LA, Heard-Costa N, Lee M, Atwood LD, for the Framingham Heart Study Investigators. 2009. Genetic Analysis Workshop 16 Problem 2: the Framingham Heart Study data. BMC Proc 3:S3.

D'Angelo GM, Rao DC, Gu CC. 2009. Combining least absolute shrinkage and selection operator (LASSO) and principal-components analysis for detection of gene-gene interactions in genome-wide association studies. BMC Proc 3:S62.

Dasarathy B. 1991. Nearest Neighbor Pattern Classification Techniques. Washington, DC: IEEE Computer Society Press.

Díaz-Uriarte R, Alvarez de Andrés S. 2006. Gene selection and classification of microarray data using random forest. BMC Bioinformatics 7:3.

Dietterich TG. 2000. Ensemble methods in machine learning. Lect Notes Comput Sci 1857:1–15.

Easton DF, Eeles RA. 2008. Genome-wide association studies in cancer. Hum Mol Genet 17:R109–R115.

Frazer KA, Murray SS, Schork NJ, Topol EJ. 2009. Human genetic variation and its contribution to complex traits. Nat Rev Genet 10:241–251.

Freund Y, Schapire R. 1997. A decision-theoretic generalization of online learning and an application to boosting. J Comput Syst Sci 55:119–139.

Friedman JH. 2001. Greedy function approximation: a gradient boosting machine. Ann Stat 29:1189–1232.

Friedman J, Hastie T, Tibshirani R. 2000. Additive logistic regression: a statistical view of boosting (with discussion). Ann Stat 28:337–407.

García-Magariños M, López-de-Ullibarri I, Cao R, Salas A. 2009. Evaluating the ability of tree-based methods and logistic regression for the detection of SNP-SNP interaction. Ann Hum Genet 73:360–369.

González-Recio O, López de Maturana E, Vega AT, Engelman CD, Broman KW. 2009. Detecting single-nucleotide polymorphism by single-nucleotide polymorphism interactions in rheumatoid arthritis using a two-step approach with machine learning and a Bayesian threshold least absolute shrinkage and selection operator (LASSO) model. BMC Proc 3:S63.

Hastie T, Tibshirani R, Friedman JH. 2001. The Elements of Statistical Learning. New York: Springer.

Heckerman D. 1999. A tutorial on learning with Bayesian networks. In: Jordan M, editor. Learning in Graphical Models. Cambridge, MA: MIT Press. p 301–354.

Hoerl AE, Kennard RW. 2000. Ridge regression: biased estimation for nonorthogonal problems. Technometrics 42:80–86.

Kim Y, Wojciechowski R, Sung H, Mathias RA, Wang L, Klein AP, Lenroot RK, Malley J, Bailey-Wilson JE. 2009. Evaluation of random forests performance for genome-wide association studies in the presence of interaction effects. BMC Proc 3:S64.

Kraja AT, Culverhouse R, Daw EW, Wu J, Van Brunt A, Province MA, Borecki IB. 2009. The Genetic Analysis Workshop 16 Problem 3: simulation of heritable longitudinal cardiovascular phenotypes based on actual genome-wide single-nucleotide polymorphisms in the Framingham Heart Study. BMC Proc 3:S4.

Lettre G, Rioux JD. 2008. Autoimmune diseases: insights from genome-wide association studies. Hum Mol Genet 17:R116–R121.

Malo N, Libiger O, Schork NJ. 2008. Accommodating linkage disequilibrium in genetic-association analyses via ridge regression. Am J Hum Genet 82:375–385.

Marchini J, Howie B, Myers S, McVean G, Donnelly P. 2007. A new multipoint method for genome-wide association studies by imputation of genotypes. Nat Genet 39:906–913.

Meier L, van de Geer S, Bühlmann P. 2008. The group lasso for logistic regression. J R Stat Soc Ser B Stat Methodol 70:53–71.

Mohlke KL, Boehnke M, Abecasis GR. 2008. Metabolic and cardiovascular traits: an abundance of recently identified common genetic variants. Hum Mol Genet 17:R102–R108.

Park T, Casella G. 2008. The Bayesian lasso. J Am Stat Assoc 103:681–686.

Pearl J. 1988. Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. San Francisco: Morgan-Kaufmann Publishers.

Plenge RM, Seielstad M, Padyukov L, Lee AT, Remmers EF, Ding B, Liew A, Khalili H, Chandrasekaran A, Davies LR, Li W, Tan AK, Bonnard C, Ong RT, Thalamuthu A, Pettersson S, Liu C, Tian C, Chen WV, Carulli JP, Beckman EM, Altshuler D, Alfredsson L, Criswell LA, Amos CI, Seldin MF, Kastner DL, Klareskog L, Gregersen PK. 2007. TRAF1-C5 as a risk locus for rheumatoid arthritis—a genomewide study. N Engl J Med 357:1199–1209.

Ripley BD. 1996. Pattern Recognition and Neural Networks. Cambridge, UK: Cambridge University Press.

Rumelhart DE, Hinton GE, Williams RJ. 1986. Learning internal representations by error propagation. In: Rumelhart DE, McClelland JL, editors. Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Cambridge, MA: MIT Press. p 318–362.

Samani NJ, Erdmann J, Hall AS, Hengstenberg C, Mangino M, Mayer B, Dixon RJ, Meitinger T, Braund P, Wichmann HE, Barrett JH, König IR, Stevens SE, Szymczak S, Tregouet DA, Iles MM, Pahlke F, Pollard H, Lieb W, Cambien F, Fischer M, Ouwehand W, Blankenberg S, Balmforth AJ, Baessler A, Ball SG, Strom TM, Braenne I, Gieger C, Deloukas P, Tobin MD, Ziegler A, Thompson JR, Schunkert H, WTCCC and the Cardiogenics

Consortium. 2007. Genomewide association analysis of coronary artery disease. N Engl J Med 357:443–453.

Schwarz DF, Szymczak S, Ziegler A, König IR. 2007. Picking single-nucleotide polymorphisms in forests. BMC Proc 1:S59.

Schwarz DF, Szymczak S, Ziegler A, König IR. 2009. Evaluation of single-nucleotide polymorphism imputation using random forests. BMC Proc 3:S65.

Sebastiani P, Ramoni MF, Nolan V, Baldwin CT, Steinberg MH. 2005. Genetic dissection and prognostic modeling of overt stroke in sickle cell anemia. Nat Genet 37:435–440.

Stassen HH, Hoffmann K, Scharfetter C. 2009. The difficulties of reproducing conventionally derived results through 500k-chip technology. BMC Proc 3:S66.

Strobl C, Zeileis A. 2008. Danger: high power!—Exploring the statistical properties of a test for random forest variable importance. Munich: Department of Statistics, University of Munich. Report No. 17.

Sun YV, Cai Z, Desai K, Lawrance R, Leff R, Jawaid A, Kardia SL, Yang H. 2007. Classification of rheumatoid arthritis status with candidate gene and genome-wide single-nucleotide polymorphisms using random forests. BMC Proc 1:S62.

Sun YV, Bielak LF, Peyser PA, Turner ST, Sheedy II PF, Boerwinkle E, Kardia SL. 2008. Application of machine learning algorithms to predict coronary artery calcification with a sibship-based design. Genet Epidemiol 32:350–360.

Sun YV, Shedden KA, Zhu J, Choi N-H, Kardia SLR. 2009. Identification of correlated genetic variants jointly associated with rheumatoid arthritis using ridge regression. BMC Proc 3:S67.

Tang R, Sinnwell JP, Li J, Rider DN, de Andrade M, Biernacka JM. 2009. Identification of genes and haplotypes that predict rheumatoid arthritis using random forests. BMC Proc 3:S68.

Tibshirani R. 1996. Regression shrinkage and selection via the lasso. J R Stat Soc Ser B Stat Methodol 58:267–288.

Tomita Y, Tomida S, Hasegawa Y, Suzuki Y, Shirakawa T, Kobayashi T, Honda H. 2004. Artificial neural network approach for selection of susceptible single nucleotide polymorphisms and construction of prediction model on childhood allergic asthma. BMC Bioinformatics 5:120.

Vapnik V. 1996. The Nature of Statistical Learning Theory. New York: Springer.

Wan X, Yang C, Yang Q, Xue H, Tang NL, Yu W. 2009. Mega-SNPHunter: a learning approach to detect disease predisposition SNPs and high level interactions in genome wide association study. BMC Bioinformatics 10:13.

Wang M, Chen X, Zhang M, Zhu W, Cho K, Zhang H. 2009. Detecting significant single-nucleotide polymorphisms in a rheumatoid arthritis study using random forests. BMC Proc 3:S69.

Wu TT, Chen YF, Hastie T, Sobel E, Lange K. 2009. Genome-wide association analysis by lasso penalized logistic regression. Bioinformatics 25:714–721.

Yang W, Gu CC. 2009. Selection of important variables by statistical learning in genome-wide association analysis. BMC Proc 3:S70.

Yuan M, Lin Y. 2006. Model selection and estimation in regression with grouped variables. J R Stat Soc Ser B Stat Methodol 68:49–67.

Ziegler A, DeStefano AL, R. KI, Bardel C, Brinza D, Bull S, Cai Z, Glaser B, Jiang W, Lee KE, Li CX, Li J, Li X, Majoram P, Meng Y, Nicodemus KK, Platt A, Schwarz DF, Shi W, Shugart YY, Stassen HH, Sun YV, Won S, Wang W, Wahba G, Zagaar UA, Zhao Z. 2007. Data mining, neural nets, trees—Problems 2 and 3 of Genetic Analysis Workshop 15. Genet Epidemiol 31:S51–S60.