BMI231 – Final Report – Paul Godden

Machine Learning for Genome-Wide Association Studies: A Critical Review

Introduction and Background

The finalisation of the Human Genome Project in April 2003, when a complete sequence was published for the first time, created the potential to identify a large set of single nucleotide polymorphisms (SNPs) across the entire genome. Consequently, this has opened the door to possibilities for great improvements in diagnosis and therapeutics (Hirschorn & Daly 2005). In the years following the Human Genome Project, genome-wide association (GWA) studies have resulted in identification of a large number of disease-susceptibility loci for several complex diseases, many of which have been successfully replicated in subsequent studies. These results have provided insight into the way these alleles relate to multifactorial traits and the resulting novel biological insights will hopefully lead to clinical advances that will change the way such diseases are managed. However, while the vast amounts of biological data hold much promise, many of the SNPs that have been discovered actually have relatively small effects upon disease susceptibility. There remains a need for improved methods of analysis in order to extract the most useful clinical information from the data (McCarthy et al. 2008). GWA studies are typically performed according to a case-control approach. Essentially, this methodology analyses the genotype from large numbers of individuals: the case group have the disease or trait under examination, while the members of the control group do not. The studies involve thousands of individuals, and examine hundreds of thousands of SNP loci – usually across the entire human genome. Each group is genotyped for a number of known SNPs and, typically, statistical analysis is performed upon the results. A typical assumption is that associated loci are in linkage disequilibrium (LD) with other variants that cause disease, or that variants occurring at particular loci are associated with changes in biological function, such as suppression of proteins known to create tumour suppression. These differing functional changes are hypothesised to result in disease susceptibility.

In addition to the fact that most SNPs discovered via GWA studies may have relatively small effects regarding such susceptibility, there are also concerns with GWA studies that have been reported in the research literature. Typically the analyses performed do not take into account any prior knowledge regarding the disease or traits in question. Further, it is often the case that only a single SNP is considered at a time, which results in linear analyses: these may be restrictive in terms of representing the complete picture of the genetic basis of a given disease or trait.

In light of this, a tendency is now occurring to look for more sophisticated analysis methods leading towards an approach that is more holistic and has greater power to explore the relationship between common sequence variations and predisposition to disease. This approach respects the complexity of the genotype–phenotype relationship, and aims to try and elucidate the genetic architecture of complex traits. Further, this methodology is focused more upon epistatic and gene–environment interactions, which have additional consequences in terms of the analytical complexity.

One of the mainstays of this new approach to analysis of GWA studies, in order to realise the benefits outlined above, is the application of machine learning (ML) methods. In this report, we review and critique some of the ML methods that have

been applied to recent GWA studies, and further discuss how effective they have been when compared with some of the standard statistical analysis methods.

Machine Learning Methods – Overview

Different ML approaches have been proposed and applied in order to model the relationship between combinations of SNPs, genetic variations and environmental factors and these relate to disease susceptibility for certain complex diseases.

Support Vector Machines

Support Vector Machines (SVMs) are applied to supervised learning problems. The key idea is to project the training data – which has relatively low dimensionality – onto a higher-dimensional feature space. In this manner, it is easier to separate, and thus distinguish, the input data. SVMs have been applied to analyse several GWA studies and are, at present, one of the more popular ML techniques for this particular application. The technique has been extremely effective in a very large number of diverse applications, and it is also noted for its computational efficiency. Further, it is possible to implement a kernel that provides a nonlinear classification boundary, which may increase the accuracy of the method for effective analysis of GWA study data.

For GWA studies, applications of SVM include Ban et al. (2010) whose study examined the importance of certain gene-gene interactions in type 2 diabetes mellitus (T2DM). In this study, 408 SNPs present in 87 genes, with evidence of association in major T2DM pathways, were analysed. There were 462 T2DM patients and 456 disease-free controls, all of Korean ethnicity. The results were promising: SVMs reported a 65.3% prediction rate, which increased to over 70% when applied to subsets of the data relating to gender. Novel associations were also discovered.

Waddell et al. (2005) applied SVM-based classification to analyse a form of cancer known as multiple myeloma. Using 3,000 SNPs in the profiling process and using a standard form of cross-validation upon the training data resulted in an accuracy of 71%. It is also noted by the authors that they used a relatively sparse set of SNPs relative to the entire genome, and thus “in future studies with a denser SNP coverage, this information would be potentially more useful.” This consideration is key to the concept of using the underlying biology to refine the particular machine learning techniques applied to the GWA study analysis.

Another study achieved good results by applying SVMs to the problem of identifying combinations of SNPs that can predict the susceptibility to breast cancer. Listgarten et al. (2004) considered the SNPs from 45 genes of potential relevance to breast cancer etiology in 174 patients as compared to the matched normal controls. They obtained an accuracy of 69% when using SVMs as the learning algorithm. The authors concluded that multiple SNPs from different genes over distant parts of the genome are better at identifying breast cancer patients than any single SNP alone.

Recently, Uhmn et al. (2009) applied several machine learning techniques including SVM to predict “patients' susceptibility to chronic hepatitis from SNPs.” Interestingly, SVMs results were slightly inferior to decision tree-based method implemented for the same data set, with classification accuracies of 67.53% and that of the decision tree is 72.68% respectively.

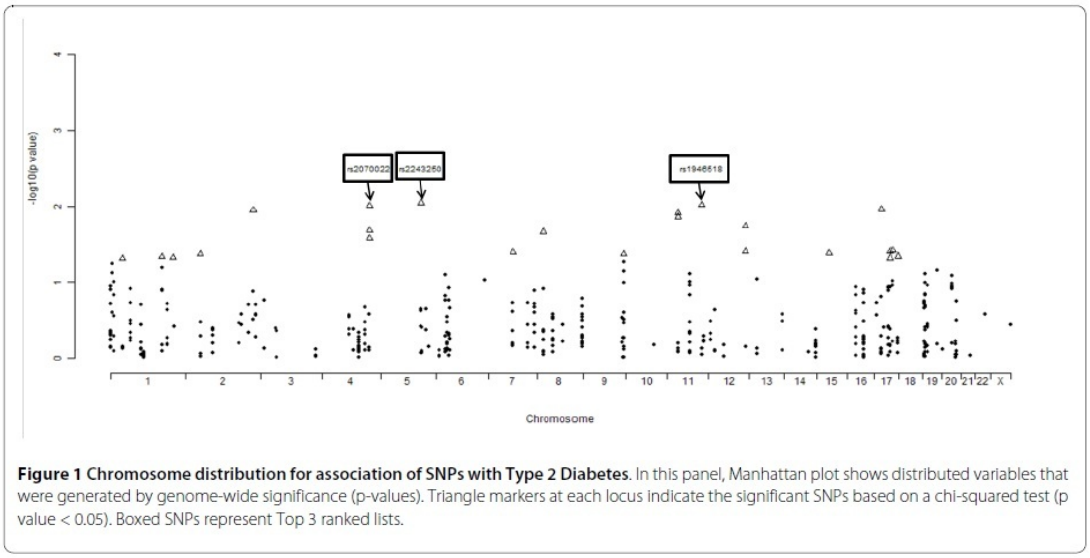

Figure 1 Chromosome distribution for association of SNPs with Type 2 Diabetes. In this panel, Manhattan plot shows distributed variables that were generated by genome-wide significance (p-values). Triangle markers at each locus indicate the significant SNPs based on a chi-squared test (p value < 0.05). Boxed SNPs represent Top 3 ranked lists.

Manhattan diagram from Ban et al. (2010). For each SNP, the p-value was calculated based upon a chi-squared test. From 408 SNPs, investigated to determine T2D susceptibility, 27 showed a significant genotype- or allele-based p-value (i.e. < 0.05).

Advantages and Disadvantages of Support Vector Machines

Unlike other ML methods, SVMs do not suffer from the 'curse of dimensionality' and there is a strong theoretical basis which guarantees a certain minimum performance (Christianni & Shawe-Taylor 2000). As outlined above, the method has discovered novel associations when applied to GWA studies and the SNP associations it has predicted are well-correlated with standard statistical analyses of the same data.

However, a different approach must be used in order to apply the method to nonlinear classification, and this increases the complexity of the implementation somewhat (Boser et al. 1992), typically by introducing radial basis functions as the kernel function. This theoretical complexity can make the method remote from the underlying genetic architecture, which can be considered a significant disadvantage despite its computational performance. It is also worth noting that the actual performance of SVMs has been very different across different GWA studies; notably, in a recent application of the method for GWA studies involving T1D and Parkinson's disease (Mittag et al.. 2012). This report is of particular interest, as the authors reported significant differences when applying the method to the two studies. For T1D, the results were excellent and commensurate with the accuracy reported elsewhere in the literature: "predictions with an area under the receiver operating characteristic curve (AUC) of ~0.88 for T1D, highlighting the strong heritable component (□90%)". However, the results were relatively poor when the same method was applied for the Parkinson's disease data, resulting in AUC ~0.56 and heritability prediction of ~38%. Further investigation via simulation studies resulted in some optimism with regard to the future effectiveness of the method with GWA studies of the latter type. However, the cautionary note does indicate that ML methods should not be expected to work "out of the box" and careful refinement to each particular study is required, at the present time.

Roshan et. al. (2011) applied both SVMs and random forests to simulated and real GWA study data sets. Their conclusions are interesting, as they show a direct comparison between two machine learning methods and a traditional statistical analysis: "We find the support vector machine to rank causal SNPs and those from associated regions higher than random forest and chi-square if applied to the top 2r chi-square-ranked SNPs, where r is the number of SNPs with p-values within

Bonferroni, and the value of r is sufficiently large." We note that the Bonferroni correction is the most stringent form of multiple hypothesis testing, and this lends extra credence to the conclusions in the analysis.

Random Forests

The use of classification and regression trees is a mainstay of machine learning methods, particularly when there is missing input data. The central idea is the creation of a graph or network-type model which, given a set of decisions (or other training data), predicts the outcome and can refine its classification given input data.

This concept can be taken a step further by creating an ensemble of many decision trees to a single set of training data, with each tree effectively 'voting' for a particular outcome, and this approach is referred to as a random forest (RF). In this case, the consensus outcome will be the final decision or prediction. We note that the RF method requires the same implementation criteria as do classification and regression trees, notably how to define the branching conditions (Biau et al. 2008). However, in principle at least, the actual predictor could be of any type: the possibility exists, for example, of a random forest of support vector machines.

Zou et al. (2012) applied RF to a GWA study involving Alzheimer's disease (AD). The results of the study, including an RF method enhanced with post-classification enrichment analysis, resulted in 1,058 SNPs with susceptibility associations being detected. Several of the resulting SNPs had previously been shown to have statistical significance with regard to AD susceptibility. Importantly, the authors noted that "the susceptible SNPs were investigated by enrichment analysis and significantly-associated gene functional annotations, such as 'alternative splicing', 'glycoprotein', and 'neuron development', were successfully discovered, indicating that these biological mechanisms play important roles in the development of AD in APOEε4 carriers". The identification of such mechanisms is a very promising direction for future research in this area, and may well present an example of an emerging trend: going beyond classification and large-scale data analysis to provide important areas for further research.

Goldstein et al. (2010) applied random forests to a GWA study concerning multiple sclerosis (MS). The case-control study involved 300,000 SNPs. One of the key findings from the research was the necessity to refine and tune the machine learning techniques in order to maximise their efficiency and eliminate noise. In the case of RF, this involved pruning based upon LD. There was found to be good agreement with the original statistical analysis of the GWA study data set, and the RF method also predicted four new candidate MS genes. It is of interest to compare this work with Zou et al. (2010), as both studies reached similarly conclusions regarding how to successfully implement RF in this context.

In another study, the RF method was applied to a GWA study investigating Crohn's disease by Schwarz et al. (2010). Each of the RF implementations analysed a simulated dataset comprising 1,006 samples genotyped at 275,153 SNPs. The four most significant SNPs detected by the RF corresponded with several GWA studies and were known to be strongly associated with susceptibility to Crohn's disease. Other novel genes that had not previously been associated were also found, and these suggest new avenues for further research.

<u>Advantages and Disadvantages of Random Forests</u>

Random forests are considered a very robust method in general, and there have been successful implementations for GWA studies, as discussed above. However, caution is advised with the method: a recent study (Nicodemus & Malley 2009)

showed that “considering correlation within predictors is crucial in making valid inferences using variable importance measures”, and this was particularly important when applying RF.

Kim et al. (2009) analysed the overall performance of RF with respect to GWA study analyses, and their conclusions summarise well the advantages and disadvantages of the method. It performed well when compared with a more traditional regression analysis; however, the authors conclude that “The causal marker that had an interactive effect with smoking did show moderate evidence of association in the RF and regression analyses, suggesting that RF may perform well at detecting such interactions in larger, more highly powered datasets.”

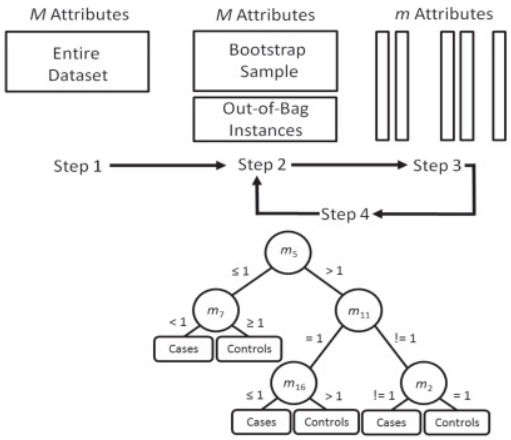

Fig. 1. Overview of the RF algorithm summarized in Section 2.3. Adapted from Reif *et al.* (2006).

An overview of the Random Forest algorithm (reproduced from Moore et al. 2010)

Multifactor Dimensionality Reduction

Multifactor Dimensionality Reduction (MDR) is a method that has been developed in order to deal with the problem of understanding high-order gene-gene interaction in GWA studies. Essentially, the method involves selection of a specific number of SNPs, and then the corresponding case-control ratios for each multi-locus genotype are calculated. Consequently, this partitioning may reveal genotypes associated with risk. Finally, cross-validation is applied in order to validate the method.

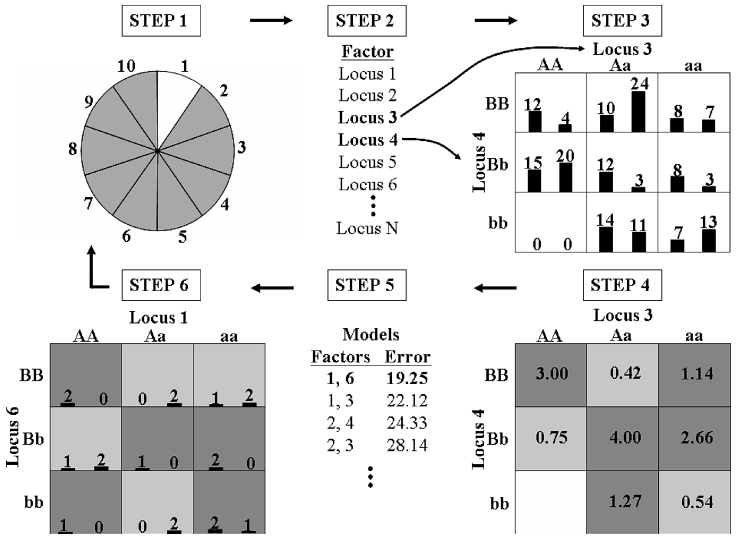

Diagram to show how MDR is applied (from Motsinger & Ritchie, 2006). In step 1, cross-validation of the data is performed. In step 2, a set of *n* genetic and / or environmental factors are selected. In step 3, the ratio of cases to controls within each multifactor class are calculated. In step four, each multifactor cell in the n-dimensional space is labelled as high risk or low risk, after comparison with a particular threshold value. In steps five and six, the model with the best misclassification error is selected and the prediction error of the model is estimated using the independent test data.

Noffisat & Motsinger-Reif (2011) applied MDR to a set of simulated GWA data in order to compare its effectiveness with both LR and other analysis methods. The aim of this research was to understand the effectiveness of MDR in epistasis; namely if this machine learning approach was able to detect gene-gene interactions in GWA studies. The approach taken was to apply three simulated types of disease models (one-locus main effect, two-locus epistatic effects, and two-locus models with joint epistatic and main effects). The authors conclude that “importantly, MDR performed as well as EC and linear regression for main effect models. It also significantly out performs LR for various two-locus epistatic models, while it has equivalent results as EC for the epistatic models. The results of this study demonstrate the potential of MDR as a filter to detect gene–gene interactions in GWAS studies.”

Another application was performed by Mahachie John et al. (2011) applied MDR to the analysis of simulated data sets. They conclude that “dealing with phenotypic mixtures and genetic heterogeneity will remain challenging for epistasis screening methods, for some time to come. Our empirical results suggest that more work is needed to better accommodate these particularities. Benefits may be gained from identifying the trait-specific factors (genetic or non-genetic) that best characterize mixed phenotypic populations.”

Ritchie et al. (2001) performed an investigation into disease susceptibility for a particular class of breast cancer. They concluded that “when it was applied to a sporadic breast cancer case-control data set, in the absence of any statistically significant independent main effects, MDR identified a statistically significant high-order interaction among four polymorphisms from three different estrogen-metabolism genes. To our knowledge, this is the first report of a four-locus interaction associated with a common complex multifactorial disease”. Although earlier than the other MDR research studies described in this report, it has the advantage of being successful when applied to experimental, rather than simulated, data.

A very recent investigation by Collins et al. (2013) applied MDR to analyse data relating to a study of tuberculosis (TB) in order to determine high-order epistatic interactions. The authors concluded: “we have identified statistically significant

evidence for a three-way epistatic interaction that is associated with susceptibility to TB. This interaction is stronger than any previously described one-way or two-way associations. This study highlights the importance of using machine learning methods that are designed to embrace, rather than ignore, the complexity of common diseases such as TB." We again note that the recent trend of ensuring that ML methods are applied with careful consideration of the fundamental biology and disease mechanism.

Advantages and Disadvantages of Multifactor Dimensionality Reduction

Implementation of MDR may complex from a computational view, as the method relies upon "search algorithms, cross-validation and permutation testing" (Moore et al. 2006). MDR has a number of advantages, however: it assumes that no hypothesis about the value of a given statistical parameter is made. Further, it assumes no particular inheritance model and is directly applicable to case-control, cohort and trio-based GWA studies. The overall predictive accuracy of the method has been good for the studies it has been applied to. However, it is limited in that it can only be used for detecting and modelling epistasis.

Naïve Bayes and Variants

The naïve Bayes classifier (NB) is a well-established machine learning method that applies Bayes theory, together with particular assumptions regarding feature independence. Several variants and improvements to the fundamental naïve Bayes implementation have been proposed for analysis of GWA study data. In this section, we briefly review some of the approaches that have been formulated and applied to recent GWA studies. One of the main concerns with application of naïve Bayes is that it may exhibit bias when there are large amounts of attributes to be analysed. Although this effect can be mitigated (Li et al. 2008), alternative formulations of the method have been proposed for GWA data in order to remedy potential issues.

In a recent investigation by Sebastiani et al. (2012) a standard naïve Bayes classifier was applied to a set of simulated data, consisting of 3,000 cases and 3,000 controls, and genotype data from 75 causal SNP and 500,000 null SNPs. The authors report good agreement between their NB approach and a traditional regression model.

Wei et al. (2011) refined the NB approach by performing model-averaging over a large number of NB models. The method was applied to a data set of late onset Alzheimer's disease in 1,411 individuals who each had 312,318 measured SNPs. The accuracy of the classifier was significantly greater than standard NB and comparable with a feature-selection NB approach. However, the model-averaged method had a run time comparable with NB, namely two orders of magnitude faster than using feature selection.

Malovini et al. (2011) applied a hierarchal Bayes model (HBM) to simulated case-control data sets with 300 elements in each set. They considered SNPs mapping to the same region of LD as "details" of the corresponding locus. Each of these details contributed to the overall effect of the region on each particular phenotype. The method was applied to both simulated data and two experimentally-acquired data sets, one for T1D and one for T2D. The former consisted of 1,963 patients affected by T1D, 1,458 control individuals from the UK Blood Service and 458,868 autosomal SNPs. The other experimental data comprised 1,924 patients affected by T2D, 1,458 control individuals and 458,868 autosomal SNPs (mapping to chromosomes 1- 22). The classification accuracy of the HNB method on both the T1D and T2D data were comparable in accuracy to other methods used to report upon the data.

One interesting approach was performed by Sambo et al. (2012), in which a variation referred to as bag of naïve Bayes (BoNB), was used. The approach is based upon NB and enhanced by three main additions: bootstrap aggregating of an ensemble of NB classifiers, a strategy for ranking and selecting the attributes used by each classifier in the ensemble and a permutation-based procedure for selecting significant biomarkers. The method was tested on case-control data for T1D derived from the WTCCC study. The BoNB method achieved "significantly higher classification accuracy" compared to a standard NB implementation and a penalised logistic regression algorithm.

Advantages and Disadvantages of Bayesian Methods

As noted earlier, Bayesian models have to be carefully applied to complex models: if a model is selected to maximise the likelihood function it may lead to overfitting of the data. However, Bayesian classifiers have been applied with success in many problems and have seen much use in bioinformatics applications. It is also possible to refine the selection of the prior distribution to ensure that there is a sufficient trade-off between the error associated with fitting the data and the complexity of the model, although this is not a trivial undertaking (Bishop, 2006).

However, NB is a very well established ML method, and relative to other ML methods it is computationally inexpensive. It has also been successfully applied in several other bioinformatics problems. The various refined Bayesian methods described above have been reasonably successful in their predictions for specific GWA studies. However, as has been seen with other ML methods, there does not currently appear to be one approach which seems significantly more accurate across multiple studies or multiple disease types, and the various approaches reviewed here remain an area of active research.

Artificial Neural Networks

Artificial neural networks (ANNs) have been applied to both supervised and unsupervised learning problems. In essence, they consist of large numbers of highly interconnected processing elements, formulated as networks that may modify their constituent weighting functions in order to adapt and respond to inputs, such as training data. ANNs were once a mainstay of machine learning techniques, but have been eclipsed in recent years as classifiers, due to the superior performance of SVMs in particular.

In terms of applications to GWA study analysis, Stassen et al. (2009) applied an ANN methodology in order to investigate the extent to which the subjects' immunoglobulin M levels can be reproducibly predicted from a multi-locus genotype. The research was based upon training data consisting of 1,042 subjects genotyped for 5,728 SNPs and a test sample of 746 subjects genotyped for 545,080 SNPs. The study showed one of the strengths of applying an ANN: namely, the ability of the method to deal with nonlinear associations. The resulting classifiers "predicted immunoglobulin M levels from the subjects' multi-locus genotypes at acceptable error rates through a configuration of 15 genomic loci (61 SNPs)."

Tomita et al. (2004) implemented a method that used an ANN implemented as follows: "For the analysis of 25 SNPs, 50 input layer units were provided. The number of hidden layer units was changed from the usual 6 to 10, to optimize the ANN for the highest possible prediction accuracy. The output layer had only 1 unit. Because the ANN model has connection weight parameters, which depend on the number of connection units, analysis of 25 SNPs with 6 hidden layer units requires 306 connection weight parameters." Their implementation was used to analyse 25

SNPs of 17 genes in a sample of 344 Japanese people, and then 10 susceptible SNPs of childhood allergic asthma were selected. The accuracy of the ANN model with 10 SNPs" was 97.7% for learning data and 74.4% for evaluation data".

Advantages and Disadvantages of Artificial Neural Networks

One disadvantage of ANNs is that when using multi-layer perceptrons, as would be expected for dealing with the types of application we are considering, care must be taken with the implementation to avoid the algorithm becoming trapped in local minima (Bishop, 2000).

Another possible disadvantage is the potential complexity of the ANN itself, which may impact the implementation of the method, with a corresponding decrease in performance of the algorithm. However, ANNs implicitly allow for nonlinear relations between the independent and dependent variables which is a great advantage when dealing with modelling multiple SNP interactions. ANNs do not require explicit distributional assumptions such as normality, and they handle missing data well (Sargent 2001). The method is particularly well-suited to problems where there is a large signal-to-noise ratio, and as seen by Tomita et al (2004) they are adaptable to studies with two and three-way interactions. However, there remain limited examples of ANNs applied to GWA study data when compared to support vector machines, for example.

Regression

Regression analysis is a well-established statistical tool for the investigation of relationships between variables. For example, in classical linear regression, the □*1, ...*□□ . The general approach has proved useful in analysis of GWA study data and there have been several investigations of its effectiveness, with refinements to the underlying method improving predictive accuracy.

Although a statistical model, there have been some notable reports that seek to meld regression and ML methods. One example is Briggs et al. (2010). The approach used was a four-stage methodology. Firstly, RF was used to identify promising regions harbouring epistatic candidates for PTPN22, a known epistatic factor, in 512 rheumatoid arthritis families consisting of 292 affected sibling pairs. The second stage was to use conventional logistic regression models to test for epistasis, assuming a multiplicative interaction. After performing a replication analysis, the final results were subject to a combined analysis from two rheumatoid arthritis data sets, with 1,624 cases and 2,506 controls. Four novel susceptibility genes were identified, and a framework for epistatic interactions resulted.

A further study (Kooperberg & Ruczinski, 2005) applied Monte Carlo logic regression (MCLR), an approach that combines Markov chain Monte Carlo and logic regression in an adaptive regression methodology. The goal was to construct predictors as Boolean combinations of binary covariates such as SNPs. The result was a collection of SNP interactions hypothesised to be associated with a disease outcome. The method was applied to a study of heart disease with 779 participants and 89 SNPs in 62 candidate genes. Comparison with statistical analyses showed results of varying accuracy, and a subsequent application of the method to a simulation study highlighted the potential for this approach with further refinements to the method.

TABLE V. Logistic regression model using predictors suggested by Monte Carlo logic regression

Predictor	Coefficient	SE	*t*-statistic	Frequency	Marginal odds ratio
1	−0.424				
$X_1 = (\text{TP53(P72R)}_d \vee \text{CBS(I278T)}_r) \wedge \text{CD14}^c_d$	0.804	0.150	5.38	48.5%	2.283
$X_2 = \text{TNFR1}_r \vee \text{APOC3(T3206G)}_d$	−1.817	0.479	−3.79	5.3%	0.153
X_3=MDM3$_d$	−0.545	0.231	−2.36	13.1%	0.599
X_4=TNFR1$_d$	−0.543	0.212	−2.57	15.8%	0.615

Monte Carlo-based regression: odds ratios from the models implemented by Kooperberg & Ruczinksi (2005).

Advantages and Disadvantages of Regression-Based Methods

There are certain disadvantages occur with the application of combined ML and regression methods. In particular, they exhibit inferior accuracy for real study data when compared with, for example, SVMs. Further, there exist differing opinions on how they should be constructed (Guan & Stephens, 2011), with no particular consensus on the optimal approach currently. However, these techniques show promise and as they continue to be developed improvements in the methodology will emerge.

Other approaches

Other methods have been applied that attempt to bridge the gap between the traditional statistical methods of GWA study analysis and the refinements provided by applying ML techniques with the relatively large training data sets available. One approach taken is to use regularisation methods to mitigate the issue of overfitting. A typical application is to apply some kind of penalty function that restricts complexity, and this is particularly useful when the data sets to be analysed are very large. One approach is to use a method referred to as Lasso - least absolute shrinkage and selection operator – which computes a particular norm for the data in question and then constrains the values of the data. This approach can effectively precondition the data and should, in principle, result in improved prediction accuracy. These methods are discussed in detail in Szymczak et al. (2009).

Due to the variety of approaches, we mention one application which is illustrative of how such methods work. Li et al. (2010) started from the assertion that we have remarked upon earlier in this report, namely that “GWAs, based on a single SNP analysis are too simple to elucidate a comprehensive picture of the genetic architecture of phenotypes. A simultaneous analysis of a large number of SNPs, although statistically challenging, especially with a small number of samples, is crucial for genetic modelling”. In order to facilitate this, they presented a framework based upon a two-stage procedure for multi-SNP modelling and analysis. The first step involved a preconditioned response variable using a supervised principle component analysis and then formulating a Bayesian-type lasso method to select a subset of significant SNPs. The Bayesian lasso was implemented via a hierarchical model, in which scale mixtures are used as prior distributions for the genetic effects and exponential priors are considered for their variances. The resulting models are then solved by using the Markov chain Monte Carlo (MCMC) algorithm. This is of interest as the potential refinements implemented by the authors are relevant to several of the other ML methods discussed above. The analysis was performed upon a data set from the FHS, a cardiovascular study based in Framingham,

Massachusetts. SNPs were chosen that could not be neglected according to the data – for example, the phenotypic data of BMI in a middle age measure of each subject in the data for a single SNP analysis. The resulting SNPs were found to be well-correlated with existing study results for this trait. The authors concluded that the use of the lasso method implied that "in this framework, SNPs with significant genetic effects can be identified more accurately."

Conclusions

We have reviewed some of the main machine learning implementations that have been applied specifically for analysis of genome-wide association studies. As the number of GWA studies continues to increase, the corresponding analysis methods will naturally need to be refined and improved in order to derive the greatest clinical and medical information from the wealth of data that GWA studies provide (Goldstein et al. 2010)

We echo the points made by Clark et al. (2004): namely, the conclusion by the authors that the on-going focus should be in two main areas. Firstly, "computational methods for data mining and machine learning", and secondly "bioinformatics methods for incorporating prior biological knowledge into data analysis algorithms". The fact that ML methods should become a mainstay of GWA study data analysis was made by Moore et al. (2010).

A recent meeting at the "Genetic Analysis Workshop 16" focused on ML approaches as "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." The review paper was published after this meeting – Szymczak et al. (2009) – suggests how some of these recommendations may be implemented, and add to these ideas below.

In conclusion, from this brief review of some of the more salient ML methods and their relative successes, we make the following comments.

1) **Ensure that nonlinearity is represented in the method.** That the underlying biology of how SNPs manifest in the genome is suggested as being nonlinear in association has been again shown in a recent GWA study focused upon Alzheimer's disease (Infant et al., 2004). Restriction to the assumption of linearity is a limitation that should be overcome, given the predictive accuracy of some of the methods discussed herein. Further, contributions such as epistasis, are not able to be analysed effectively within such a linear framework (Marchini et al. 2005).

2) **Use as much metadata as possible**. With such vast amounts of data available, it is possible to envisage GWA studies that have close association with established haplotypes – for example, the international HapMap project (http://snp.cshl.org/index.html). The ability for machine learning methods to closely integrate their results with such databases may be very powerful. This may also allow the relaxation of the fact that most existing GWA study analyses are "single-SNP" analyses, which simply test each SNP, one at a time, for association with the phenotype.

3) **Use the underlying biology.** One of the most key aspects that machine learning methods should adopt is to be guided by the underlying biology and the corresponding allelic architecture. It is expected that as the methods become more widespread for GWA studies, and thus the applications become more refined, they

will be tailored according to the underlying biology. This is partly covered by the assumption of nonlinearity, discussed above. One notable example is HaploBuild (Laramie et al., 2007) which presents an alternative method to haplotype creation than the 'sliding window' approach that is frequently used. One key effect of this approach is that when neighbouring SNPs are in strong LD, the windowing method may not capture appropriate haplotype diversity. Although not a machine learning method per se, HaploBuild indicates that respect of the underlying biology can result in methods with greater utility. Another example of an approach that does not use machine learning is Holmans et al. (2013). In Pirooznia et al. (2012) the authors conclude that "A common challenge for all the classifiers we tested is the problem of appropriate feature selection. Although we found that the prediction improved with increasing numbers of SNPs included in the classifier models, it is conceivable that we could do even better if we were able to identify and include only the most etiologically relevant sets of SNPs". Genotype-phenotype associations would be one outcome of such an approach.

4) **Use the identification of novel SNPs to guide research focus upon disease** Disease association and identification of traits suggest relevance to a particular disease or condition are key objectives for GWA studies. However, even producing suggestions for further investigation, in the case of novel SNP associations has shown to be of great value. One example was the discovery of an allele in PNPLA3 (Romeo et al., 2008), and the consequent understanding of this allele is involved in hepatic triglyceride metabolism. Machine learning, appropriately implemented, can assist with finding such associations and thus suggesting avenues for further research.

   One method that we suggest to implement many of the above points is to use an entirely novel approach, borrowed from a different field, for machine learning applied to analyse GWA studies. Our approach is based upon a promising method that has been used to predict earthquakes (Oh et al., 2008). The nature of the problem is quite analogous: relatively large sets of data which require classification based upon factors which have a certain threshold significance. The key point here is that the method is based upon Bayesian analysis which, as we have already seen, has had some success when applied to GWA studies, but this particular refinement includes automatic relevance detection. This consideration has not yet been implemented and it would be of interest to investigate further if such an approach is viable in the current context.

   We conclude by mentioning the exciting possibilities promised by GWA studies, but also by the emergence of powerful and effective machine learning techniques to unlock the great biomedical potential of this area of research.

References

1) Hirschorn JN & Daly MJ, "Genome-wide association studies for common diseases and complex traits", Nat Rev Genet. 2005 6(2)

2) McCarthy MI et al, "Genome-wide association studies for complex traits: consensus, uncertainty and challenges", Nat Rev Genet. 2008 9(5)

3) Ban et al. "Identification of Type 2 Diabetes-associated combination of SNPs using Support Vector Machine" BMC Genetics 2010, 11:26

4) Waddel N. et al. “Predicting cancer susceptibility from single-nucleotide polymorphism data: a case study in multiple myeloma”, BIOKDD '05 Proceedings of the 5th international workshop on Bioinformatics 21 - 28 ACM New York, 2005

5) Listgarten J, Damaraju S, Poulin B, Cook L, Dufour J, Driga A, Mackey J, Wishart D, Greiner R, Zanke B: “Predictive models for breast cancer susceptibility from multiple single nucleotide polymorphism” Clin Cancer Res 2004, 10:2725-2737

6) Uhmn S, Kim D-H, Ko Y-W, Cho S, Cheong J, Kim J, ”A study on application of single nucleotide polymorphism and machine learning techniques to diagnosis of chronic hepatitis” Expert Systems 2009, 26:60-69

7) Christianni N & Shawe-Taylor J, “An Introduction to Support Vector Machines and Other Kernel-based Learning Methods”, CUP 2000

8) Boser, BE, Guyon, IM, Vapnik, VN. “A training algorithm for optimal margin classifiers” 5th Annual ACM Workshop on COLT, 144–152, ACM Press 1992.

9) Mittag F, et al. “Use of support vector machines for disease risk prediction in genome-wide association studies: concerns and opportunities“, Hum Mutat. 2012 33(12)

10) Roshan U et al., “Ranking causal variants and associated regions in genome-wide association studies by the support vector machine and random forest”, Nucleic Acids Research, 2011, 39(9)

11) Biau G et al. “Consistency of Random Forests and Other Averaging Classifiers”, Journal of Machine Learning Research 9 (2008) 2015-2033

12) Zoul L et al. “A genome-wide association study of Alzheimer's disease using random forests and enrichment analysis.” Sci China Life Sci. 2012 55(7)

13) Goldstein et al. “An application of Random Forests to a genome-wide association dataset: Methodological considerations & new findings” BMC Genetics 2010, 11:49

14) Schwarz DF et al. “On safari to Random Jungle: a fast implementation of Random Forests for high-dimensional data” Bioinformatics, 26(14) 2010

15) Nicodemus KK & Malley JD, “Predictor Correlation Impacts Machine Learning Algorithms: Implications for Genomic Studies”, Bioinformatics. 2009, 25(15)

16) Kim Y et al. “Evaluation of random forests performance for genome-wide association studies in the presence of interaction effects” BMC Proceedings 2009, 3(7)

17) Motsinger AA & Ritchie MD “Multifactor dimensionality reduction: An analysis strategy for modelling and detecting gene–gene interactions in human genetics and pharmacogenomics studies” Human Genomics, 2006 2(5) 318–328

18) Noffisat OO and Motsinger-Reif AA “Multifactor Dimensionality Reduction as a Filter-Based Approach for Genome Wide Association Studies” Front. Genet., 2011 2(80)

19) Mahachie John JM et al. “Model-Based Multifactor Dimensionality Reduction to detect epistasis for quantitative traits in the presence of error-free and noisy data”, European Journal of Human Genetics 2011, 19(6)

20) Ritchie MD et al. “Multifactor-Dimensionality Reduction Reveals High-Order Interactions among Estrogen-Metabolism Genes in Sporadic Breast Cancer”, Am. J. Hum. Genet. 2001, 69:138–147

21) Collins RL et al. “Multifactor dimensionality reduction reveals a three-locus epistatic interaction associated with susceptibility to pulmonary tuberculosis”, BioData Mining 2013, 6(4)

22) Moore JH et al. “A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility“, Journal of Theoretical Biology (2006), 241:252–261

23) Li L, et al. “A Method for Avoiding Bias from Feature Selection with Application to Naive Bayes Classication Models“ Bayesian Analysis (2008), 3(1)

24) Sebastiani P et al. “Naïve Bayesian classifier and genetic risk score for genetic risk prediction of a categorical trait: not so different after all!” Front Genet. 2012; 3(26)

25) Wei W et al. “The application of naive Bayes model averaging to predict Alzheimer’s disease from genome-wide data” J Am Med Inform Assoc 2011 18(3)

26) Malovini et al. “Hierarchical Naïve Bayes for genetic association Studies”, BMC Bioinformatics 2011, 13(14)

27) Sambo F et al. “Bag of Naïve Bayes: biomarker selection and classification from genome-wide SNP data” BMC Bioinformatics 2012, 13(14)

28) Bishop, CM “Pattern recognition and machine learning”, Springer, 2006

29) Stassen, HH et al. “The difficulties of reproducing conventionally derived results through 500k-chip technology”, BMC Proceedings 2009, 3(7)

30) Tomita Y et al., “Artificial neural network approach for selection of susceptible single nucleotide polymorphisms and construction of prediction model on childhood allergic asthma” BMC Bioinformatics 2004, 5:120

31) Bishop CM, “Neural Networks for Pattern Recognition”, Oxford University Press, 2000

32) Sargent DJ. “Comparison of artificial neural networks with other statistical approaches. Results from medical data sets”. Cancer. 2001, 91:1636–42

33) Briggs FBS et al. “Supervised machine learning and logistic regression identifies novel epistatic risk factors with PTPN22 for rheumatoid arthritis”, Genes Immun. 2010, 11(3).

34) Kooperberg C, Ruczinski I, “Identifying Interacting SNPs Using Monte Carlo Logic Regression”, Genetic Epidemiology 28: 157–170 (2005)

35) Guan, Y and Stephens, M “Bayesian Variable Selection Regression For Genome-Wide Association Studies And Other Large-Scale Problems”, The Annals of Applied Statistics 2011, 5(3), 1780–1815

36) Szymczak S et al. “Machine Learning in Genome-Wide Association Studies”, Genetic Epidemiology 2009, 33(1)

37) Li, J et al. “The Bayesian Lasso for Genome-wide Association Studies” Bioinformatics 2011, 27(4)

38) Clark et al. Clark,A.G. et al. (2004) “Determinants of the success of whole-genome association testing.”, Genome Res. 2004, 15:1463–1467

39) Goldstein BA et al. “Application of Random Forests to a genome-wide association dataset: Methodological considerations & new findings” BMC Genetics 2010, 11:49

40) Infante J et al. “Gene-gene interaction between interleukin-1A and interleukin-8 increases Alzheimer’s disease risk” J. Neurol. 2004, 251:482–483

41) Marchini J et al. “Genome-wide strategies for detecting multiple loci that influence complex diseases” Nature Genetics 2005, 37(4).

42) Laramie, JL et al. “HaploBuild: an algorithm to construct non-contiguous associated haplotypes in family based genetic studies” Bioinformatics. 2007, 23(16)

43) Holmans P et al. “A pathway-based analysis provides additional support for an immune-related genetic susceptibility to Parkinson's disease” Hum Mol Genet. 2013, 22(5)

44) Pirooznia M et al. “Data Mining Approaches for Genome-Wide Association of Mood Disorders” Psychiatr Genet. 2012, 22(2)

45) Romeo S et al. “Genetic variation in PNPLA3 confers susceptibility to nonalcoholic fatty liver disease” Nat Genet. 2008, 40(12)

46) Oh, CK et al. “Bayesian Learning Using Automatic Relevance Determination Prior with an Application to Earthquake Early Warning”, Journal of Engineering Mechanics, 134(12), 2008