Livestock Science 166 (2014) 217–231

Contents lists available at ScienceDirect

Livestock Science

journal homepage: www.elsevier.com/locate/livsci

Machine learning methods and predictive ability metrics for genome-wide prediction of complex traits ☆

Oscar González-Recio [a,b,*], Guilherme J.M. Rosa [c,e], Daniel Gianola [c,d,e]

[a] Biosciences Research Division, Department of Environment and Primary Industries, Agribio, 5 Ring Road, Bundoora, VIC 3083, Australia
[b] Dairy Futures Cooperative Research Centre, Bundoora, VIC 3083, Australia
[c] Department of Animal Sciences, University of Wisconsin-Madison, WI 53706, USA
[d] Department of Dairy Science, University of Wisconsin-Madison, WI 53706, USA
[e] Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, WI 53706, USA

ARTICLE INFO

Keywords:
Animal breeding
Cross validation
Genome wide prediction
Machine learning
Nonparametric
Predictive accuracy

ABSTRACT

Genome-wide prediction of complex traits has become increasingly important in animal and plant breeding, and is receiving increasing attention in human genetics. Most common approaches are whole-genome regression models where phenotypes are regressed on thousands of markers concurrently, applying different prior distributions to marker effects. While use of shrinkage or regularization in SNP regression models has delivered improvements in predictive ability in genome-based evaluations, serious over-fitting problems may be encountered as the ratio between markers and available phenotypes continues increasing. Machine learning is an alternative approach for prediction and classification, capable of dealing with the dimensionality problem in a computationally flexible manner. In this article we provide an overview of non-parametric and machine learning methods used in genome wide prediction, discuss their similarities as well as their relationship to some well-known parametric approaches. Although the most suitable method is usually case dependent, we suggest the use of support vector machines and random forests for classification problems, whereas Reproducing Kernel Hilbert Spaces regression and boosting may suit better regression problems, with the former having the more consistently higher predictive ability. Neural Networks may suffer from over-fitting and may be too computationally demanded when the number of neurons is large.

We further discuss on the metrics used to evaluate predictive ability in model comparison under cross-validation from a genomic selection point of view. We suggest use of predictive mean squared error as a main but not only metric for model comparison. Visual tools may greatly assist on the choice of the most accurate model.

1. Introduction

The availability of technology for assaying thousands of genomic variants simultaneously in a cost-effective way has revolutionized the paradigm of prediction of genetic merit or phenotypes of individuals as well as animal and plant breeding strategies (Meuwissen et al., 2001).

☆ This paper is part of the special issue entitled: Genomics Applied to Livestock Production, Guest Edited by Jose Bento Sterman Ferraz.

* Corresponding author at: Biosciences Research Division, Department of Environment and Primary Industries, Agribio, 5 Ring Road, Bundoora, VIC 3083, Australia.
E-mail addresses: oscar.gonzalez-recio@depi.vic.gov.au (O. González-Recio), grosa@wisc.edu (G.J.M. Rosa), gianola@ansci.wisc.edu (D. Gianola).

http://dx.doi.org/10.1016/j.livsci.2014.05.036
1871-1413/

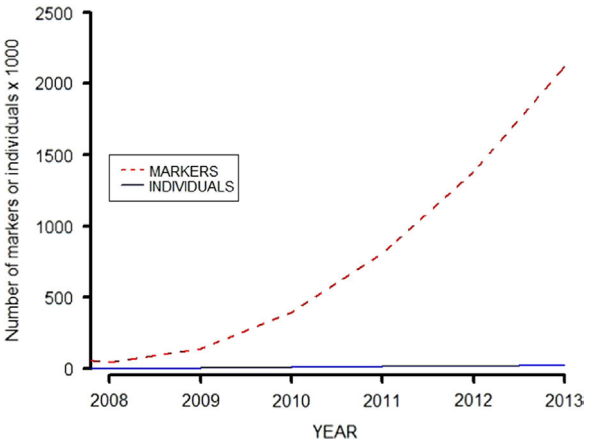

Fig. 1. Relationship between the number of marker effects to be estimated in regression models and the number of genotyped individuals with phenotype in livestock populations.

New statistical approaches have been developed for dealing with the "large *p* small *n* problem" (*p*: number of markers, *n:* sample size) because, typically, only a few hundred or thousands of individuals are genotyped whereas the number of molecular markers is much larger, and it is constantly increasing due to technological changes (Fig. 1). For instance, in 2008 and 2009 the number of markers available in single nucleotide polymorphism (SNP) chips for livestock species (e.g., bovine, swine, sheep) was around 50,000 and the largest data set represented between 1000 and 4000 individuals genotyped (Gonzalez-Recio et al., 2008; Moser et al., 2009; VanRaden and Sullivan, 2010). Hence, at that time, the *p:n* ratio was between 12.5 and 50. Around 2010, some higher density chips appeared (e.g., BovineHD beadchip of Illumina), covering 777,000 markers whereas the larger consortia in dairy cattle (North American consortium and Eurogenomics) included around 16,000 genotyped animals with phenotypes, which represents a ratio of 48.5 marker effects to be estimated per phenotype available in the sample (Lund et al., 2011).

Use of sequencing technologies places further challenges because several million of variants per individual may need to be taken into account in predictive models. Although the larger international consortiums are projected to expand to reach around 30,000 individuals with phenotypes, this is yet a *p:n* ratio of 80–100. Thus, the dimensionality problem is becoming more and more acute in recent years, with major implications at the level of statistical inference (Gianola, 2013).

While use of shrinkage or regularization in SNP regression models, e.g., Bayes A, Bayes B, Bayes C, BayesCpi, Bayesian Lasso, Bayes R (see Gianola, 2013 for a discussion), has delivered improvements in predictive ability in genome-based evaluations (de los Campos et al., 2013), serious over-fitting problems may be encountered where *p:n* exceeds 50–100. Models using genomic similarity matrices between individuals alleviate the dimensionality problem by capturing 'overall' signals while still exploiting multi-locus linkage disequilibrium between the unknown (and elusive) QTLs and the markers (Habier et al., 2007).

Machine learning is an alternative approach for prediction and classification, capable of dealing with the dimensionality problem in a flexible manner. Bayesian regression methods cope with regularization by assigning different prior distributions (purportedly aiming to reflect "genetic architecture") to marker effects. However, machine learning approaches provide a larger and more general suite of flexible methods for this purpose. Several studies using Reproducing Kernel Hilbert Spaces Regression (RKHS), Neural Networks (NN), Support Vector Machines (SVM), Random Forests (RF) and Boosting have been used for genome-enabled prediction in livestock and plants (González-Recio and Forni, 2011; Long et al., 2011a, 2011b; Ober et al., 2011; Vazquez et al., 2012; González-Recio et al., 2013; Crossa et al., 2014). This article provides an overview of non-parametric methods that have been applied to genome-wide prediction (GWP) in animal breeding so far. Subsequently, we outline two non-parametric alternatives for coping with the high dimensionality problem arising in sequence-assisted evaluation. Finally, we discuss pros and cons of metrics used for assessing predictive ability in cross-validation of models implemented in genomic evaluation, and suggest non-parametric alternatives for this purpose. This paper does not address design of cross-validations, which have been reviewed elsewhere (e.g. Hastie et al., 2009, 241 pp.).

2. Non-parametric models used in genome-assisted prediction

Most machine learning approaches to genome assisted evaluation are based on regressing phenotypes on some function of SNP genotype codes $g(\mathbf{x})$, as

$$\mathbf{y} = \mathbf{1}\mu + \begin{bmatrix} g_1(\mathbf{x}_1) \\ g_2(\mathbf{x}_2) \\ \dots \\ g_n(\mathbf{x}_n) \end{bmatrix} + \mathbf{e}$$

This function $g_i(\mathbf{x}_i)$ (i denotes individual and $\mathbf{x}_i$ is the observed p-dimensional genotype of i) is assumed to be an approximation to the true genetic merit of each individual, after adjusting phenotypes for environmental effects and expressing the latter as a deviations from the population mean, μ. Then, $\mathbf{1}$ is a column vector of ones. The correction for nuisances can be done model-wise, but we assume the data have been pre-corrected in some reasonable manner, for simplicity. The vector $\mathbf{e}$ represents residuals, typically assumed $N(0, \mathbf{I}\sigma_e^2)$, where $\mathbf{I}$ is an $n \times n$ identity matrix and σ_e^2 is a residual variance.

When $g(\mathbf{x}_i) = \sum_{j=1}^{p} x_{ij}\beta_j$, where x_{ij} is a code for the genotype at marker j on individual i, the model reduces to a linear regression, considered here as a parametric approach, typically because regression coefficients have an interpretation in the light of additive gene action theory. A recent review of linear models applied in GWP is in de los Campos et al. (2013). Machine learning methods, including non-parametric regression, do not assume linear and additive action of markers a priori, but the type of function given by $g(\mathbf{x})$ determines the learning attained. Here, we will describe and discuss functions defining a RKHS, SVM, boosting, RF and NN. The first two methods are based on a single function with good learning ability (i.e., a strong learner), whereas the others can be viewed as combinations