Extreme Learning Machine: A New Approach for Genomic Prediction of Complex Traits

A. Ehret[1], **D. Hochstuhl**[2] **and G. Thaller**[1]

[1]Institute of Animal Breeding and Husbandry, University Kiel, Germany
[2]Institute for Theoretical Physics and Astrophysics, University Kiel, Germany

ABSTRACT: A wide range of methods for predicting phenotypes based on genomic data has become available. Increasingly, the focus is also being set to machine learning methodology. Despite this progress, the prediction of complex traits from high density SNP-panels remains an extremely demanding task, particularly in terms of the computational effort required in calibration and prediction. We present a fast learning algorithm for artificial neural networks which was introduced by Huang et al. in 2004. Our experimental results show that this approach is able to achieve good generalization performance with much less computational effort while outperforming the traditional gradient-based learning in artificial neural networks, which is a great advantage when analyzing high dimensional data. We demonstrate the capabilities of the new approach to genomic predictions in animal and plant breeding.

Keywords: genomic selection; extreme learning machine; complex traits

Introduction

Today, in animal and plant breeding, non- and semi-parametric, as well as machine learning methods, have become increasingly important in performing genome-enabled predictions of complex traits. At the current times of high-density panels and sequence information, prediction become challenging in terms of computational costs and efficiency. The number of markers now vastly exceeds the number of records, with the effect that typical fitting methods, such as least squares regression, often require some prior variable selection or shrinkage estimation procedure. When using machine learning approaches, one is usually facing a large number of parameters, non-linear training procedures and the problem of over-fitting. Therefore, the computational costs are usually even larger than those of standard parametric methods, which only require the solution of linear systems. A notorious example is the back-fitting algorithm for the training of artificial neural networks, which must be processed iteratively. However, the additional effort is usually compensated by an improved generalization performance. To overcome this computational challenge, a fast learning algorithm for single-hidden-layer feed-forward neural networks, called extreme learning machine (ELM), was proposed by Huang and coworkers (Huang et al. (2004), Huang et al. (2006)).

Since its invention, the ELM has had a tremendous impact in the field of machine learning as well as on its close relatives, such as data mining and pattern recognition. For a survey on its various applications see Huang et al. (2011). The reason for its success is twofold. On one hand, the underlying algorithms are almost trivial when compared to more sophisticated methods such as support vector machines (SVM) or artificial neural networks trained by back-propagation (to which we refer here as ANN). On the other hand, despite this simplicity, several studies have revealed that its predictive performance is generally comparable to that of the standard methods mentioned (SVM and ANN), with often diminished training times. In this contribution, we employ the ELM to predict milk traits from dairy cattle data with molecular marker information, which is to our knowledge the first application of this kind, and compare the results to an ANN approach with the same input and output configuration.

Materials and Methods

Extreme Learning Machine. The theory of the Extreme Leaning Machine may be approached from two directions, starting either from ANN, or from basis function expansion methods. As the neural network perspective seems to be common in the literature, we consider the second alternative in the brief summary that follows. For this, assume we are given a data set $(x_i, y_i)_{1\le i\le N}$ of size N, where $x_i \in \mathbb{R}^M$ denotes an M-dimensional predictor and $y_i \in \mathbb{R}$ is the corresponding response variable. A common method to describe the relationship between the predictors and the observations is to assume a linear expansion into a number of K basis functions $f_k\colon \mathbb{R}^M \to \mathbb{R}$,

$$E(y|x;ß) = F(x;\beta) = \sum_{k=1}^{K} \beta_k f_k(x) \qquad (1)$$

Given this model, the task is then to minimize the residual sum of squares of the data set,

$$\min_{\boldsymbol{\beta}} \sum_{j=1}^{N} \left(y_j - F\left(x_j;\beta_1,\dots,\beta_K\right)\right)^2 \qquad (2)$$

which, through variation with respect to β, quickly leads to the usually over-determined (since $K \le N$) linear system

$$\begin{pmatrix} f_1(x_1) & \cdots & f_K(x_1) \\ \vdots & \ddots & \vdots \\ f_1(x_N) & \cdots & f_K(x_N) \end{pmatrix} \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_K \end{pmatrix} = \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix} \qquad (3)$$

Denoting the components by $\boldsymbol{F}$, $\boldsymbol{\beta}$, and $\boldsymbol{y}$, the least-squares solution of this equation is given by

$$\boldsymbol{\beta} = \boldsymbol{F}^{\dagger}\boldsymbol{y} \qquad (4)$$

where $\boldsymbol{F}^{\dagger}$ stands for the Moore-Penrose pseudo-inverse (Golub and van Loan (2012)). Though there are several methods to calculate this pseudo-inverse, the singular value decomposition is usually a convenient choice due to its good stability properties (Press (2007)). Solving the least-squares problem for the basis expansion model therefore requires a SVD of a matrix of size $N \times K$, which for $K \leq N$ scales as $\mathcal{O}(KN^2)$.

Note that the predictor dimension M has not entered the derivation so far, and will eventually do so only in the function evaluation step for the construction of the coefficient matrix. Thus, at least in principle, this allows for large dimensions of the predictor space to be treated, as is the case in this work.

The crucial point in model (1) is the selection of an appropriate basis set, and several statistical methods exist which differ only in this respect, such as linear regression, logistic regression, spline interpolation, and many more. The particular choice leading to the ELM is given by

$$f_k(x) = g\left(\sum_{m=1}^{M} w_{km} x_m + b_k\right) \qquad (5)$$

Here, w_{km} and b_k are randomly chosen real numbers (according to any random distribution), whereas g denotes a function which in the ELM is usually taken to be the sigmoid function.

In combination with the model (1), equation (5) allows for its interpretation as a single-hidden layer artificial neural network with randomly chosen connections and biases. More precisely, this network consists of M input neurons, K hidden neurons with sigmoid activation function and a single output neuron with linear activation function. w_{km} stands for the connection strength between input neuron m and hidden neuron k, b_k for the corresponding bias (intercept), and the coefficients β_k are the connection strengths between the k-th hidden node and the output layer, as shown in Figure 1.

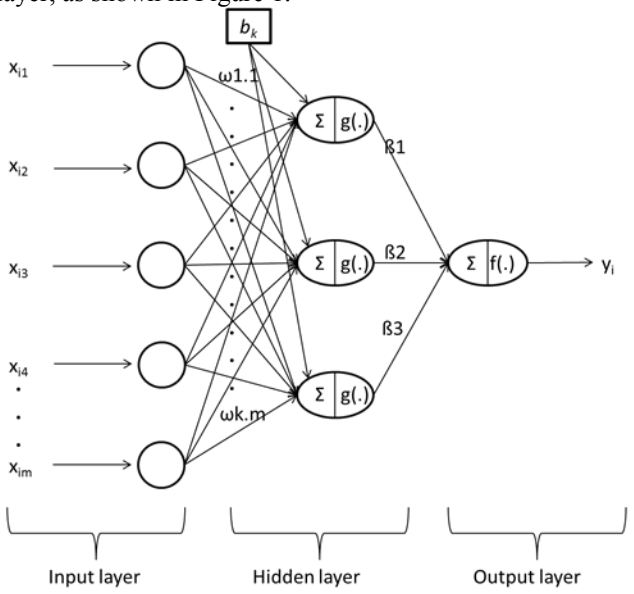

Figure 1. ELM architecture of a Single-hidden layer artificial neural network

With the model being the same as a conventional neural network, the major difference that makes up the ELM is its training. The standard method proceeds by optimizing all involved parameters, i.e., w_{km}, b_k and β_k. Due to their non-trivial dependencies, the training needs to be done layer by layer in a process called back-propagation. In contrast, the ELM randomly fixes a large number of coefficients in advance, and leaves only the task to determine the output layer connection strengths. Due to the linear activation function in the output node, this can be accomplished by the least-squares solution outlined in Eqs. (2) to (4). Further, the method presented can easily be extended to include a regularization parameter (via modification of Eq. (2)) or to use kernels such as radial Gaussian functions (via modification of (5)).

Benchmark model. To compare the predictive ability of the new algorithm with a standard network we used a single-hidden layer artificial neural network with back-propagation algorithm as benchmark. Here, the connections and biases were chosen optimally using a given training set. To avoid poor generalization ability early stopping was implemented in training. Both models, ANN and ELM, were set up with a sigmoid activation function in the hidden layer and a linear function in the output layer. For the ANN, an architecture of ten hidden neurons achieved the best predictive ability when dealing with the entire marker set.

Data. Genotypic and phenotypic information of 3,341 German Fleckvieh bulls were employed in the machine. All animals were genotyped with a 50k SNP-panel and after quality control 39,344 SNP markers remained in the analyses. Quality control included the elimination of SNPs with minor allele frequency < 0.05 and missing genotype frequency > 0.95. For the remaining loci, missing genotypes were imputed using the population based imputing algorithm Minimac (Howie et al. (2012)), and haplotypes were inferred using MaCH (Li et al. (2010)).
DYD of three milk traits (milk yield, fat yield and protein yield) were used as phenotype records in the analysis. To account for ANN and ELM peculiarities feature scaling was applied to both phenotypic and genomic data, so the data was normalized to the [-1,1] range, to enhance numerical stability.

To assess the predictive performance of the ELM and of the standard ANN, a 10 fold cross-validation scheme was used. The dataset was randomly split into ten equal folds of phenotypes and genotypes. Then, nine folds were used for training and one for testing. This was rotated ten times so that every fold served as test set once. The average correlations between predicted and true phenotype in the testing sets of all runs then were used to evaluate predictive ability.

Results and Discussion

The results represent our first and recent application of the ELM and serve as an illustration of the advantages of this promising approach. Results are given in Table 1, showing the correlation obtained in the cross-validation runs for a ELM with 1000 hidden nodes (the number of hidden nodes needs to be much larger than for the ANN, in order to compensate for the random

initialization). The ELM was able to give predictions with accuracy well comparable to that of the ANN approach. The crucial point here is that ELM needed significantly less computation time. On average, for a ten-fold cross-validation, the ELM required 1.3 hours, whereas the ANN with ten neurons in the hidden layer needs 11.2 hours for the same task. Another advantage is that, with increasing numbers of hidden neurons, the prediction performance of the ELM was prone to increase as well. As an example, we applied the cross-validation procedure to an ELM with 10000 hidden neurons for the trait protein yield. This resulted in a largely increased average correlation coefficient of 0.618 between true and predicted phenotype (computation time for a ten-fold cross-validation: 26,6 hours). As shown in our recent work (Ehret et al. to be published), it is much more difficult to tune the standard ANN to such levels of prediction accuracy, when the entire genotype matrix is used for predictions.

Table 1. Estimated predictive abilities for three milk traits

Trait	ELM	ANN
	r	r
Milk yield	0.468	0.469
Fat yield	0.453	0.453
Protein yield	0.450	0.477

r= average Person correlation of cross-validation runs, ELM= Extreme learning machine with 1000 hidden neurons, ANN= Artificial neural network with 10 hidden neurons and back-propagation

Conclusion

With a low number of neurons the ELM matches the performance of the ANN in predicting milk traits from genomic information, while using the whole marker information but drastically reducing the computational cost. Increasing the number of hidden neurons produced promising results, as shown for the trait protein yield. Although the study was based on limited data, our results suggest that the ELM is likely to be useful for predicting complex traits using high-dimensional genomic information, a situation where the number of coefficients that need to be estimated exceeds sample size. The ELM, in the same way as SVM or ANN, has the ability of capturing nonlinearities, but its great advantage is in keeping the computational cost at a reasonable level. The predictive ability seemed to be enhanced by using a larger number of neurons in the ELM, which is in contrast to the standard ANN approach. However, further studies are needed to confirm these results and explore the capabilities of the ELM.

In summary, the ELM is a promising method for prediction of future phenotypes from high dimensional genomic data sets, as has been suggested by the presented study. Hence, a detailed investigation of its capabilities will be an important component in our future work.

Literature Cited

Golub, G. H., and Van Loan, C. F. (2012). Matrix computations (Vol. 3). JHU Press.

Huang, G. B., Zhu, Q. Y., and Siew, C. K. (2004). Neural Networks Proceedings 2004 on IEEE International Joint Conference, volume 2:985-990.

Huang, G. B., Zhu, Q. Y., and Siew, C. K. (2006). Neurocomputing, 70: 489-501.

Huang, G. B., Wang, D. H., and Lan, Y. (2011). International Journal of Machine Learning and Cybernetics, 2:107-122.

Howie, B., Fuchsberger, C., Stephens, M., et al. (2012). Nature genetics, 44: 955-959.

Li, Y., Willer, C. J., Ding, J., Scheet, P. and Abecasis, G. R. (2010). Genetic epidemiology, 34: 816-834.

Press, W. H. (2007). Numerical recipes 3rd edition: The art of scientific computing. Cambridge university press.