Semi-Supervised Learning Combining Phenotyped and Non-phenotyped individuals for Enhancing Prediction in Residual Feed Intake

***C. Yao*[1], X. Zhu[2] and K.A. Weigel[1]**
[1]Department of Dairy Science University of Wisconsin, Madison, WI, USA, [2]Department of Computer Science University of Wisconsin, Madison, WI, USA

ABSTRACT: Genomic prediction is challenging for residual feed intake (RFI), because the costly measurement on individual feed intake limits the size of reference population. To improve the genomic prediction accuracy in RFI, we introduced self-training model (one of semi-supervised learning strategies), as a novel method combining phenotyped and non-phenotyped individuals. It trained the model using its own predictions on non-phenotyped animals. The results suggested that self-training wrapped around support vector machine increased the prediction accuracy up to 3% using about 1,000 non-phenotyped animals. The improvement increased as the number of non-phenotyped animals included increased, but may approach a plateau. This method can be particularly helpful for enhancing the genomic prediction on new traits such as RFI at the early stage, when the size of reference population is limited. The extension to other traits needs to be further studied.

Keywords: genomic prediction; semi-supervised learning; residual feed intake; dairy cattle

Introduction

Genomic prediction has been successfully making faster progress than traditional progeny testing in dairy cattle. To perform functional predictions, a sufficient number of animals are usually needed to construct the reference population. It is especially challenging for new traits such as feed efficiency, where a massive reference population has not been available. The main reason is that individual feed intakes must be measured, which not only is difficult but also costs exorbitantly. However, with several years of genomic selection on production traits, hundreds of thousands of bulls and cows have been genotyped. There can be much more non-phenotyped animals than phenotyped animals available without any extra cost. The challenge is that how we can benefit genomic prediction from non-phenotyped animals.

One powerful tool from machine learning community to address this question is semi-supervised learning. As the name suggests, it is a type of models between unsupervised and supervised learning. Current popular genomic prediction models mostly fall into the category of supervised learning, such as GBLUP and Bayesian families. When doing genomic prediction, phenotype (such as milk yield) is provided as the desired label on an animal to supervise the training (i.e., learning) process based on its genotype. Whereas in unsupervised learning, no phenotype is available to supervise how the animal should be handled. One example of unsupervised learning tasks is principal component analysis for clustering, where the goal is to separate animals into groups based on the variation from genotypes.

As one of the widely used semi-supervised learning algorithms, self-training model was used in our study to enhance genomic prediction on residual feed intake (RFI) measuring feed efficiency. It has had many remarkable applications in genetics, such as increasing the accuracy of gene start prediction by combining models of protein-coding and non-coding regions and models of regulatory sites near gene start (Besemer et al. (2001)) and gene finding in eukaryotic genomes by estimating directly from yet anonymous genomic DNA (Lomsadze et al. (2005)). Intuitively, self-training model extends supervised learning to include additional information from unsupervised learning. During the learning process, the model uses its own predictions to teach itself. In other words, we extended one typical genomic prediction model (i.e., support vector machine (SVM) in this study), by including both cows measured on individual feed intake (phenotyped animals) and cows that did not have individual feed intake available (non-phenotyped animals) in the training set. It allowed the model to be trained using its own predictions on non-phenotyped animals.

The objective of this study was to train a genomic prediction model on RFI from both phenotyped and non-phenotyped animals, such that it predicted better than the model trained on the phenotyped animals alone.

Materials and Methods

Animals and data collection. Phenotyped animals were 792 lactating Holstein and crossbred dairy cows in the L. C. Allenstein Dairy Herd at the University of Wisconsin – Madison, and had feed intake and production data measured from 50 to 200 days in milk. Animal use procedures were approved by the Animal Care and Use Committee of the College of Agricultural and Life Sciences at the University of Wisconsin-Madison. Outlier records were identified to eliminate potential errors from equipment or measurement error. The expected ranges of milk yield, milk composition, dry matter intake (DMI), and body weight (BW) were set as the population means plus and minus 5 standard deviations (SD). Records beyond this range were trimmed to the boundary values. Because very few cows had DMI measurements from multiple lactations, only one lactation per animal was considered. Weekly means were calculated for milk yield, milk composition, DMI, and BW. Missing weekly mean milk composition and BW were set to the mean values of the previous and following weeks. Non-phenotype animals were 1,127 Holstein cows from 4 research herds (University of Florida,

Iowa State University, Michigan State University, and Virginia Tech University).

Residual feed intake. In this study, RFI was defined as the deviation of an animal's feed intake from the average intake of its cohort, after adjustment for milk production, maintenance, and known environmental differences. The weekly mean RFI of 792 phenotyped animals were the residual terms calculated as:

$$DMI = \mu + YSC + ParAge + dim + NEL + MBW + ration + RFI$$

where DMI was a vector of weekly average DMI, μ was the population mean, YSC was the year-by-season at calving, ParAge was the parity-by-age at calving, dim was the average days in milk of the week, MBW was the metabolic BW (i.e., BW to the 0.75 power), and ration was the random cohort effect. Then, the RFI of each animal equaled the mean of all weekly mean RFI.

Genetic markers. All genotypes of 644 (out of 792) phenotyped animals and 1127 non-phenotyped animals were obtained from the USDA–ARS Animal Improvement Programs Laboratory (Beltsville, MD). Genotypes of cows on low density SNP chip were imputed to higher density using genotype information from over 562,000 bulls and cows in the database. Missing SNP genotypes were filled in with rounded allele frequencies in the current US Holstein population as of February 2014. The SNP genotype at each locus was coded as 0, 1, or 2, counting the number of minor allele copies. The SNPs with minor allele frequencies less than 5% were removed. A total of 57,491 SNPs per individual for both phenotyped and non-phenotyped animals were available for genomic prediction analysis.

Semi-supervised learning. Self-training model as one of semi-supervised learning strategies was used in this study. Phenotyped animals were separated into the training and testing sets based on birthdate. The training set comprised 540 cows with genotype (G_1) and phenotype (P_1) having birthdate before Jan. 1, 2010, whereas the testing set included 104 cows with genotype (G_T) and phenotype (P_T) born after Jan. 1, 2010. The only contributions from non-phenotype animals were their genotypes (G_2). The "svm" function from "e1071" package Version 1.6-1 in R (http://cran.r-project.org/web/packages/e1071/index.html) was used for SVM with radial basis kernel and default parameters tuned within the training set. The algorithm is described below, and also illustrated in Figure 1.

Algorithm 1. Self-training.

Step 1: *Train a SVM predictor f from phenotyped animals using G_1 and P_1.*

Step 2: *Predict phenotype ($\hat{P}_2$) for non-phenotyped animals based on G_2.*

Step 3: *Add G_2 and $\hat{P}_2$ to G_1 and P_1, and train another SVM predictor, f^*.*

In the testing phase, compare accuracies of f and f^* on the testing set (R_{SL} and R_{SSL}). Firstly, predict phenotypes $\hat{P}_T$ and $\hat{P}_T^*$ based on G_T using f and f^*, respectively. Secondly, calculate R_{SL} (R_{SSL}) as the correlation between $\hat{P}_T$ ($\hat{P}_T^*$) and P_T.

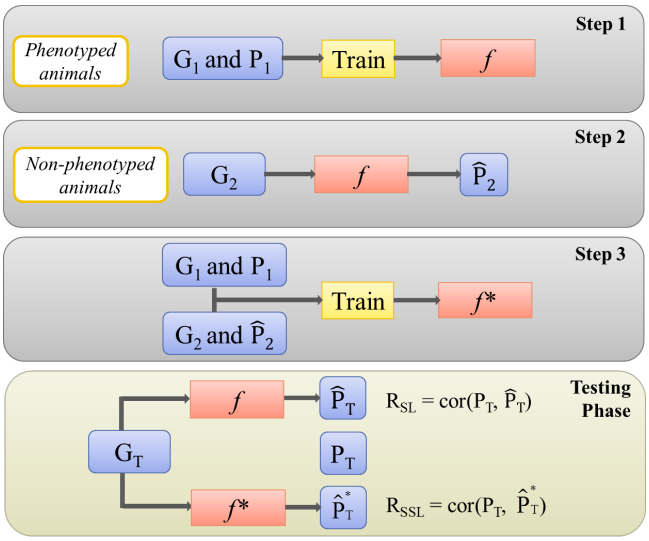

Figure 1: The illustration of self-training algorithm. Step 1: train a SVM predictor f from phenotyped animals using G_1 and P_1. Step 2: predict phenotype ($\hat{P}_2$) for non-phenotyped animals based on G_2. Step 3: Add G_2 and $\hat{P}_2$ to G_1 and P_1, and train another predictor, f^*. In the testing phase, compare accuracies of f and f^* on the testing set (R_{SL} and R_{SSL}). Firstly, predict phenotypes $\hat{P}_T$ and $\hat{P}_T^*$ based on G_T using f and f^*, respectively. Secondly, calculate R_{SL} (R_{SSL}) as the correlation between $\hat{P}_T$ ($\hat{P}_T^*$) and P_T.

Results and Discussion

The first aim was to access how the presence of non-phenotyped animals affected the prediction performance. Figure 2 shows the progress of prediction accuracy for the fixed number of phenotyped animals (540 cows) and with different number of non-phenotyped animals. The results were for 10 different samples of non-phenotyped animals for each number of 200, 400, ..., 1000. The accuracy of fully-supervised SVM for the same number of 540 phenotyped animals (R_{SL} = 29.2%) is also plotted. The results suggested that including non-phenotyped animals into the predictor improved the prediction accuracy, and the improvement increased as the number of non-phenotyped animals included increased.

Secondly, we tested how the improvement of prediction accuracy changed with different number of phenotyped animals used (Figure 3). The improvement was calculated by subtracting the accuracy of the initial supervised learning from the final accuracy of semi-supervised learning, i.e. $R_{SSL} - R_{SL}$. The results were averaged for 10 different samplings of phenotyped animal subsets with size of 300, 400, and 500 over non-phenotyped

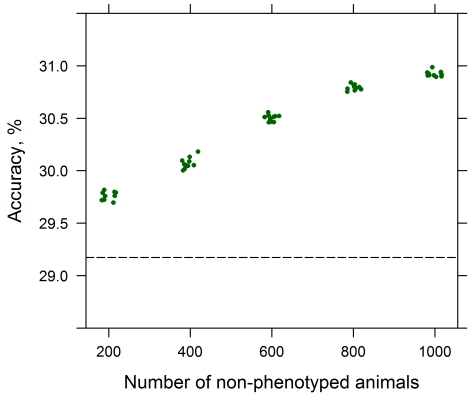

Figure 2: Accuracies of fully-supervised genomic prediction trained on 540 phenotyped animals (long dash) and accuracies of predictions using self-training after adding 200, 400, 600, 800 and 1,000 non-phenotyped animals (green dot).

animals of 200, 400, ..., 1000. The plots indicated that with the same number of non-phentyped animals, the improvement obtained depended on the number of phenotyped animals used. The benefit of adding the fixed number of non-phenotyped animals was larger for the model that included smaller number of phenotyped animals. A similar outcome was also reported by Filipovych et al. (2010) when using semi-supervised leaning to classify brain images of patients with uncertain diagnoses. The reason is that semi-supervised learning, in general, is to address the scarcity of phenotype animals. When the number of phenotyped animals was sufficient, the information that may be added from non-phenotyped animals would be limited. Therefore, this method can be particularly helpful for enhancing the genomic prediction on RFI and possibly other new traits at the early stage when the reference population is small.

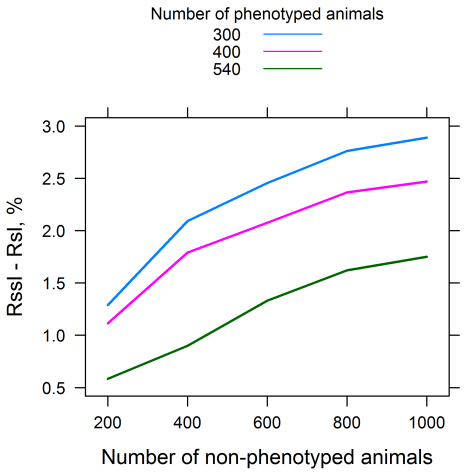

Figure 3: The improvement of genomic prediction accuracy ($R_{SSL} - R_{SL}$) started with 300, 400 and 500 phenotyped animals and different number of non-phenotyped animals.

We also observed that slopes of all four curves decreased as the number of non-phenotyped animals increased. In other words, the improvement in prediction accuracy associated with an increase from 200 to 400 non-phenotyped animals was greater than the improvement in accuracy from 400 to 600, and so on. With this trend, we may approach a plateau in the improvement of accuracy in terms of including more non-phenotyped animals. Hence, starting with a fixed number of phenotyped animals, there may be a maximum amount of accuracy that can be improved by adding information from non-phenotyped animals using self-training.

The major advantages of self-training are its simplicity and the fact that it can use any prediction model. For example, the model can also be GBLUP or Bayesian families. The self-training only wraps around the predictor without change its inner workings. However, due to the fact that we let the model learn from its own predictions, an early mistake made by the model can reinforce itself and lead to a worse predictor in the next step. It can happen if assumptions of the model were not appropriate on the data, or if the number of phenotyped animals was not large enough to training a perfect predictor to start with in the step 1. Various heuristics have been introduced to address this problem (Zhu and Goldberg (2009)). In fact, as noted by many researchers, the semi-supervised learning does not always help (Elworthy (1994); Cozman et al. (2003)), and therefore the extensions of self-training on traits other than RFI needs to be further studied.

Conclusion

We introduced self-training model (one of semi-supervised learning strategies) as a novel method to improve the genomic prediction accuracy of residual feed intake. The results suggested that self-training wrapped around SVM increased the prediction accuracy by adding genomic information from non-phenotyped animals. The improvement increased as the number of non-phenotyped animals included increased, but may approach a plateau. This method can be particularly helpful for enhancing the accuracy of genomic prediction on new traits such as RFI at the early stage when the reference population is small, but the extension to other traits needs to be further studied.

Literature Cited

Besemer, J., Lomsadze, A., and Borodovsky, M. (2001). Nucleic Acids Res., 29:2607-2618.

Cozman, F., Cohen, I. and Cirelo, M. (2003). In ICML-03, 20th International Conference on Machine Learning., 99-106.

Elworthy, D. (1994). In Proceedings of the 4th Conference on Applied Natural Language Processing, 53–58

Filipovych, R., and Davatzikos, C. (2010). NeuroImage, 55:1109–1119.

Lomsadze, A., Ter-Hovhannisyan, V., Chernoff, Y.O. et al. (2005). Nucleic Acids Res., 33:6494-6506.

Zhu, X. and Goldberg, A. (2009). *Morgan and Claypool Publishers.*