Proc. Assoc. Advmt. Anim. Breed. Genet. 25: 154 - 157

COMPARING GENOMIC PREDICTION ACCURACIES FOR COMMERCIAL COWS' REPRODUCTIVE PERFORMANCE USING GA2CAT AND TWO MACHINE-LEARNING METHODS

Y. Li[1], S. Hu[1], L. Porto-Neto[1], R. McCulloch[1], S. McWilliam[1], J. McDonald[2], C. Smith[2], P. Alexandre[1], S. Lehnert[1] and A. Reverter[1]

[1]CSIRO Agriculture & Food, St Lucia, QLD, 4067 Australia
[2]MDH Pty Ltd, Cloncurry, QLD, 4824 Australia

SUMMARY

Heifers' second joining pregnancy and lactation status (PLS) is an important fertility trait for commercial cattle herds in North Queensland. Genomic prediction of a candidate bull's contribution to its female progeny's PLS presents a technical challenge because the trait has a non-ordinal multi-class nature. We previously developed a new algorithm, Genomic Attributions to a Categorical Trait (GA2CAT) to tackle the problem. However, the merit of the method has not been evaluated against those of machine learning methods. In this study, using two commercial cow populations (795 and 340 cows respectively) with high-density SNP genotypes and imbalanced PLS phenotypes, we compared the classification performance of the new method GA2CAT with two machine learning approaches (Random Forests (RF) and Support Vector Machines (SVM)). The results from a five-fold cross-validation scheme indicate that the classification accuracy of GA2CAT was greatly impacted by the coding system of PLS categories. For highly imbalanced non-ordinal multiclass datasets, using the average overall accuracy value for evaluating the classification performance of the GA2CAT and ML methods was misleading and Matthews correlation coefficient values should be applied.

INTRODUCTION

Female reproductive traits directly impact the profitability of commercial beef herds. Among many reproductive traits, fertility-related ones are the most important. In dairy and beef cattle, they are measured by a range of continuous (e.g. age of puberty, days at first calving), binary (e.g. pregnancy status) or count traits (e.g. number of inseminations) (Toghiani *et al*. 2017). However, in Australian northern commercial cattle herds, following natural syndicate joining, heifers are usually mustered and grouped based on the result of their 2nd joining pregnancy and lactation status (PLS). Females can be assigned to six PLS categories: 1. DNP = Dry and Not Pregnant; 2.WNP = Wet and Not Pregnant; 3. DEP = Dry and Early Pregnant; 4. DMP = Dry and Mid Pregnant; 5. DLP = Dry and Late Pregnant; 6. WEP = Wet and Early Pregnant (Reverter *et al.*2016). This non-ordinal multi-class phenotype presents a technical challenge when trying to rank potential sires based on their genomic relationships with phenotyped heifers. To address this issue, we have developed a new method called Genomic Attributions to a Categorical Trait (GA2CAT) to predict an individual sire's contribution to its future daughters' performance (Li *et al.* 2022). However, the performance of GA2CAT has not been benchmarked against other methods commonly used for analysing non-ordinal multi-class traits, such as the machine learning (ML) based Random Forests (RF) and Support Vector Machines (SVM). Therefore, we conducted the study to compare genomic prediction accuracies of GA2CAT and two ML methods.

MATERIALS AND METHODS

Datasets. Two datasets containing 1,135 tropical Brahman cows, 795 from the 2020 season (referred to as Cows_795) and 340 from the 2021 season (Cows_340), from a north Queensland commercial property were used for the study. All animals with PLS records were individually

genotyped for 54,791 SNPs (Neogen Australasia GGP TropBeef 50K chip) which were then imputed to high density using 700K genotypes of 861 legacy BeefCRC Brahman cattle as the reference genome. Table 1 summarises the composition of the phenotype records in both populations, illustrating unevenly distributed multi-class categories.

Phenotypic data recoding. For comparison purposes, three different phenotype recording systems for PLS records were investigated (Table 1). These include: a) treating PLS as a binary trait (2PLS, Non-pregnant "1" vs pregnant "2"); b) as a four-category trait (4PLS, Dry and Non-Pregnant "1", Wet and Non-Pregnant "2", Dry and Pregnant "3", and Wet and Pregnant "4"); and c) as a six-category trait (6PLS, see Table 1 for details).

Table 1. Composition of 2nd Joining Pregnancy and Lactation Status (PLS) records of two Brahman cow populations (795 and 340 cows respectively) and three phenotype recording systems

		Cow population		Phenotype recoding system		
PLS	Code	Cows_795	Cows_340	2PLS*	4PLS*	6PLS*
Dry and Non-Pregnant	DNP	124	61	1	1	1
Wet and Non-Pregnant	WNP	358	109	1	2	2
Dry and Early Pregnant	DEP	77	109	2	3	3
Dry and Mid Pregnant	DMP	70	45	2	3	4
Dry and Late Pregnant	DLP	86	6	2	3	5
Wet and Early Pregnant	WEP	80	10	2	4	6
Total		795	340			

*2PLS: binary categories, 4PLS: four categories; 6PLS: 6 categories

Statistical methods. Three analytical methods were used for evaluating classification accuracy, including GA2CAT (Li *et al*. 2022), RF (Berriman 2001) and SVM (James *et al.* 2013). In brief, the GA2CAT algorithm applies a standard genomic relationship matrix derived from the method of VanRaden (2008) between the reference and testing populations to predict the likely contributions of an individual animal in the testing population to individual classes of a categorical trait. For PLS, a GA2CAT value of a given animal for a given PLS category is defined as the animal's average genomic relationship with other animals having that PLS category divided by its average genomic relationship across all animals. RF is based on ensemble learning of a large number of decision trees deriving from random sampling of various subsets (both SNPs and animals) of a given dataset. It takes the average of decision trees (with replacement) to improve the predicted accuracy of the dataset. The final output (variable importance value) of RF is based on the majority votes of predictions. SVM applies different kernel functions (linear or non-linear) to identify a hyperplane that maximizes the separation of the data points to their potential classes (binary or multi-classes). While a genomic relationship matrix was used for deriving the GA2CAT values, both RF and SVM directly applied SNP genotypes for the analyses.

A 5-fold cross-validation scheme was used for evaluating the classification performance of each method. Each cow population was randomly divided into 5 equal-size groups and each group (68 in Cows_340 or 159 animals in Cows_795) was in turn used as the validation set. Overall accuracy ((true positive + true negative)/(true positive + true negative + false positive + false negative)) was used for evaluating the prediction performance. The final results were based on the average prediction accuracy of five validation groups. Given the imbalanced multiclass datasets used here, we also applied the Matthews correlation coefficient (MCC, Chicco and Jurman 2020) as a measure of the quality for multiclass classification. MCC values normally range from -1 to 1, with 1 representing a perfect prediction, 0 an average random prediction, and -1 a perfect misprediction.

Hyperparameter tuning for RF and SVM. A range of hyper-parameter values was examined for each ML method to determine the critical parameters that minimize prediction errors. These include: for RF, the size of forest trees (Ntree =100, 500), and the number of SNP markers at each sampling event (Mtry = 100, 500, 1000 and 5000); for SVM, insensitivity zone (gamma = 0.001, 1, 5, 10) and the penalty parameter (C= 0.001, 1, 10). All other parameters for each method took default values. The RF and SVM classifiers in the "scikit-learn" Python package (Pedregosa *et al*. 2011) were used for classification predictions.

RESULTS AND DISCUSSION

Comparison of classification performance of GA2CAT, RF and SVM. The overall average prediction accuracies (standard deviations in the brackets) of the three methods from a five-fold cross-validation scheme are summarised in the top part of Table 2. When changing the coding of PLS from two to four to six categories, the overall classification accuracy decreased significantly in both populations for all methods in the small population Cows_340, but to a much lesser extent in the large population Cows_795.

Table 2. Classification performance of GA2CAT, RF and SVM under different PLS coding systems in two cow populations, using a five-fold cross-validation scheme. A) The overall average classification accuracies (standard deviations in brackets); b) Matthews correlation coefficients (MCC)

	Cows_340			**Cows_795**		
Method	GA2CAT	RF	SVM	GA2CAT	RF	SVM
A. Overall Accuracy			Cow population			
2PLS	0.46 (0.097)	0.51 (0.073)	0.47 (0.032)	0.53 (0.027)	0.61 (0.029)	0.61 (0.033)
4PLS	0.18 (0.034)	0.43 (0.063)	0.47 (0.018)	0.24 (0.027)	0.44 (0.061)	0.45 (0.052)
6PLS	0.091 (0.024)	0.25 (0.054)	0.29 (0.034)	0.12 (0.025)	0.46 (0.052)	0.45 (0.052)
B. MCC			Cow population			
2PLS	-0.071 (0.19)	0.020 (0.143)	0.000 (0.000)	0.059 (0.049)	0.077 (0.040)	0.000 (0.000)
4PLS	-0.039 (0.029)	-0.037 (0.036)	0.000 (0.000)	0.013 (0.026)	-0.027 (0.043)	0.000 (0.000)
6PLS	-0.017 (0.036)	-0.062 (0.073)	0.000 (0.000)	-0.023 (0.036)	0.053 (0.043)	0.000 (0.000)

RF: Random Forest; SVM: Support Vector Machine.

The poor performance of the three methods under 6PLS could be due to the phenotype of PLS being a non-ordinal multi-class categorical trait. The separation of animals for three Dry and Pregnant classes, i.e. early, mid, and late pregnancy was not as clean-cut as those in the binary situation (2PLS, non-pregnant vs pregnant). For the GA2CAT, the genomic relationships between animals in these three classes in the training populations were very similar, therefore the predicted contributions of the animals in the validation populations to six categories of PLS (i.e. GA2CAT values) were very similar. As a result, it made the correct assignment of the animals in the testing

populations to different categories extremely difficult. The results indicate the necessity of recoding PLS records before applying different analytical methods to achieve reliable results.

Across two cow populations, for the same coding system, e.g. 6 categories (6PLS), the two ML methods (RF and SVM) seemed to outperform the GA2CAT (see the average accuracies in Table 2). The margin was large in the population Cows_795 (0.46 (RF), 0.45 (SVM) vs 0.12 (GA2CAT). The difference between RF and SVM was little in comparison to either of them with the GA2CAT. However, when investigating further on the classes correctly classified, we found that both RF and SVM assigned all of the individuals in the validation datasets to the category of Wet and Non-Pregnant. This was the class with the largest number of phenotypic observations in Cows_795. This confirms the downside of ML methods that bias toward the majority class by over-sampling the abundant classes and under-sampling minor classes (Chicco and Jurman 2020).

When evaluating the performance of three methods by the MCC values (the lower half of Table 2), all three methods had the MCC values either zero (SVM) or close to zero. These suggest that: a) the phenotype PLS is a low heritability trait, as all three methods followed a random prediction behavior (MCC values ~ 0.00). In addition, the accuracy values for the GA2CAT fitted the random sampling expected prediction accuracies of 0.5 (PLS2), 0.34 (PLS4) and 0.25 (PLS6); b) there was no significant classification performance difference among the GA2CAT, RF and SVM.

CONCLUSION

The results from a five-fold cross-validation scheme indicate that different coding systems of PLS categories greatly impacted the classification outcome of the GA2CAT. For highly imbalanced non-ordinal multiclass datasets, using the average overall accuracy value for evaluating the classification performance of the GA2CAT and ML methods was misleading and MCC values should be applied. A GA2CAT value is the weighted average of genomic relationships between reference and validation populations for a particular category, it reflects better the heritable nature of a phenotypic trait.

ACKNOWLEDGEMENT

The authors wish to acknowledge Meat and Livestock Australia (MLA) for their financial support to the project P.PSH.1211 “Validation of pooled DNA gEBV for Brahman commercial cow fertility”.

REFERENCES

Breiman L. (2001) *Mach. Learn*. **45:** 5.

Chicco D.,and Jurman G. (2020) *BMC genomics.* **21:** 1.

James G., Witten D., Hastie T. and Tobshirani (2013) ‘An Introduction to Statistical Learning’. Springer, Heidelberg, Germany.

Li Y., Lehnert S.A., Porto-Neto L., McCulloch R., McWilliam S., Alexandre P, McDonald J., Smith C. and Reverter A. (2022) Proceedings of 12th World Congress on Genetics Applied to Livestock Production. Rotterdam. The Netherlands. 3-8 July 2022.

Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondel M., Prettenhofer P., Weiss R., Dubourg V., Vanderplas J., Passos A., Cournapeau D., Brucher M., Perrot M. and Duchesnay É. (2011). *J. Mach Learn Res* **12:** 2825.

Reverter A., Porto-Neto L.R., Fortes M.R., McCulloch R., Lyons R.E., Moore S., Nicol D., Henshall J. and Lehnert S.A (2016) *J Anim Sci*. **94:** 4096.

VanRaden P.M (2008). *J. Dairy Sci*. **91:** 4414.