Imputation of genetic characteristics using deep learning methods

D. Segelke, L. Gehrke and J. Wabbersen

Vereinigte Informationssysteme Tierhaltung w.V. (vit), Verden/Germany

Abstract

Different imputation methods are used to deal with missing markers and to infer genetic characteristics. In routine genetic evaluation, the majority of adopted imputation methods are pedigree and population based. In this study, we compare the routinely used methods with innovative methods based on deep learning and other machine learning frameworks. Therefore, the frameworks Keras, LightGBM and a combination of these methods are compared to the common software tool Beagle. Imputation accuracy for four different genetic characteristics were analysed. Results show that a combination of Keras and LightGBM outperform Beagle significantly in accuracy and the computation time decreases drastically. The results also demonstrate that big datasets and the presence of close related animals in the training set are needed. In conclusion, machine learning methods, such as deep learning, are novel powerful tools, which can improve the efficiency of breeding programs.

Key words: deep learning, keras, SNP, imputation, genomic prediction

Introduction

In routine genetic evaluation, different chip densities are extrapolated to a common sizes by imputation (Segelke et al., 2012). Moreover, imputation is used to infer genetic characteristics (Segelke et al., 2013). Imputation can be done by pedigree based, (VanRaden et al., 2013 or Sargolzaei et al., 2014) or population based imputation methods (e.g. Beagle: Browning & Browning, 2007). In times of artificial intelligence, deep learning and other machine learning methods become more popular. Especially for image analysis these methods produce outstanding results in contrast to linear models. Aim of the present study was to compare the performance of machine learning methods to routinely used imputation methods using the example of four different genetic traits.

Materials and Methods

Cholesterol deficiency (CD; Kipp et al., 2015), HH3 (McClure et al. 2013), Kappa Casein (Caroli et al., 2009) and polledness (Rothammer et al. 2014) were exemplary chosen for this study. The four traits are routinely analysed with the current EuroGenomics medium density chip and the previous EuroGenomics low density SNP chips. Table 1 gives an overview of the number of animals in the training and validation set per genetic characteristics. Training animals were born between 2011 and 2018. All validation animals were born in 2019. Minor allele frequency varied between validation and training data set (Table 1) because the validation animals were younger and selection for valuable traits like the kappa casein B-Allele and polledness selected against HH3 and CD, respectively.

Table 1. Number of animals and minor allele frequencies (MAF) for different genetic characteristics

	CD	Kappa Casein	HH3	Polled
No. of training animals	242,600	428,974	406,867	406,250
No. of validation animals	33,292	33,873	33,275	33,289
MAF training (%)	2.4	34.8	2.5	4.9
MAF validation (%)	1.8	39.9	1.9	7.1

Missing 50K marker on the LD chip were imputed with FImpute V2 (Sargolzaei et al., 2014).

The software package Beagle (v. 1398; Browning & Browning, 2007) was compared to the deep learning framework Keras (Francois et al., 2015) and the gradient boosting framework LightGBM (LGBM; Ke et al., 2017). Beagle is a genetic optimized algorithm using population based information. 20 imputations and phasing iteration were chosen. The deep learning library Keras is used with the Tensorflow (Abadi et al., 2015) backend. An example of the chosen model and number of hidden layers and outputs can be found in figure 1. LGBM is a fast gradient boosting framework that uses decision trees.

Parameters chosen are, i.a, a learning rate of 0.02 and a maximum of 20,000 boosted trees with early stopping. In addition to the individual predictions of the neural network and the gradient boosting machine, the average of the two predicted probabilities is used as another prediction (Ensemble).

Accuracy was measured by the correlation between imputed and true genotypes.

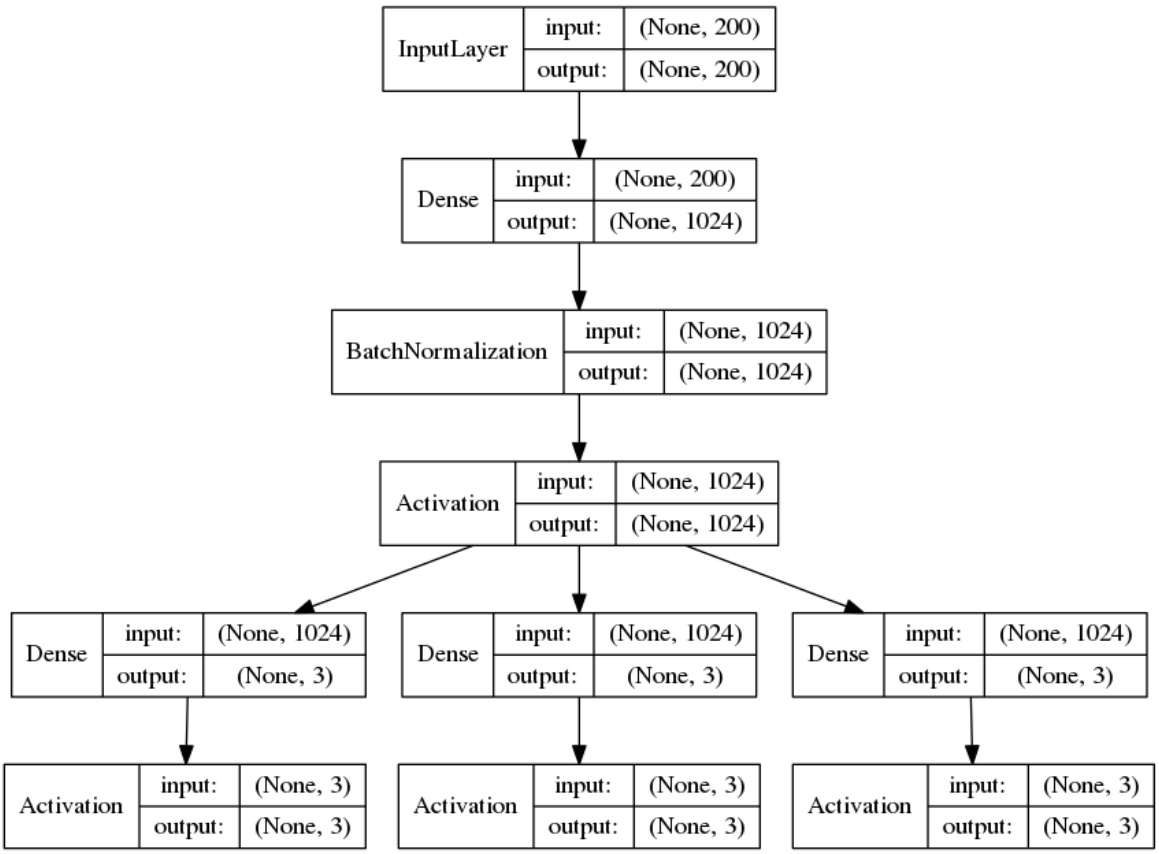

Figure 1. Keras model plot for polled

Results

Table 2 shows the imputation accuracy per genetic characteristic. For all traits the ensemble method gave the highest imputation accuracy. In contrast, Beagle had the lowest accuracies for all traits and the highest computation time varying between three to eight hours. All other methods were extremely fast with a computation time between one to eight minutes.

Table 2. Imputation accuracy for different traits and methods

Trait	Beagle	Keras	LGBM	Ensemble
Polled	98.7	98.8	98.8	98.9
CD	94.9	96.5	96.7	97.1
HH3	98.9	99.1	99.4	99.5
Kappa Casein	99.6	99.6	99.6	99.6

Table 3 illustrates the accuracy of validation animals by their relationship to the reference population for polledness. Imputation accuracy for Beagle clearly depends on the presence of relatives in the training population although Beagle does not use the pedigree information as input variable. Keras shows a similar pattern. For LGBM no clear trend can be found, here the scenarios, in which only dam or only sire was presented in the training population had higher accuracies than the two other scenarios.

Table 3. Accuracy of validation by their relationship to the reference population (polledness)

Presence of relatives in training population	Beagle	Keras	LGBM
Sire & dam (n=10,035)	99.3	99.6	99.0
Only dam (n=16,764)	99.2	99.3	99.2
Only sire (n=19,136)	99.1	99.4	99.0
Neither sire nor dam (n=16,525)	98.4	98.3	98.6

Discussion

Deep learning and machine learning methods need big data sets to outperform classical methods. To analyses this effect different sizes of training datasets were simulated (5,000, 25,000, 50,000, 100,000, 200,000, and 400,000). The results show, that for using data sets with less than 400,000 animals, Beagle had better imputation results in terms of accuracy compared to the other models. However, using the biggest datasets, all three methods outperform Beagle.

Conclusion

Imputation accuracy can be improved by using deep learning or other machine learning algorithms instead of Beagle. The computation time decreased drastically by using the innovative frameworks. The combination of the LGBM and the Keras model results in the highest accuracy, but large datasets are needed to outperform existing methods. Close relatives in training population are important for all frameworks.

Outlook

Due to the linearity of breeding values, the advantages of deep and machine learning frameworks might be limited for genetic and genomic evaluation. Nonetheless, those frameworks provide great potential to improve various branches of animal breeding industry. For example, for the routine imputation from lower density to a common size. First results show that missing SNPs on LD chips can be imputed to a common reference size using deep learning methods with a similar accuracy as FImpute.

Furthermore, the deviation of new phenotypes by sensor data or images can be analyzed using deep learning. MIR spectral data analysis and data anomaly detection (plausibility checks) can be improved.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C. Corrado, G. S., Davis, A., Dean, J. Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jozefowicz, R., Jia, Y., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Schuster, M., Monga, R., Moore, S., Murray, D., Olah, C., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y. and Zheng, X. 2015. TensorFlow: Large-scale machine learning on heterogeneous systems, Software available from tensorflow.org.

Browning, S., R. and Browning B., L. 2007. Rapid and accurate haplotype phasing and missing data inference for whole genome association studies by use of localized haplotype clustering. Am J Hum Genet 81:1084-97.

Caroli, A. M., S. Chessa and G. J. Erhardt. 2009. Invited review: Milk protein polymorphisms in cattle: Effect on animal breeding and human nutrition. J. Dairy Sci. 92:5335-5352.

Francois, C. et al. Keras. 2015. https://keras.io

Ke, G., Meng, Q. Finley, T. Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T. 2017. LightGBM: A Highly Efficient Gradient Boosting. Decision Tree. 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

Kipp, S., Segelke, D., Schierenbeck, S., Reinhardt, F., Reents, R., Wurmser, C., Pausch, H., Fries, R., Thaller, G., Tetens, J., Pott, J., Piechotta, M. and Grünberg, W. 2015. A new Holstein haplotype affecting calf survival. Interbull Bull. 49:49–53.

McClure, M., E. Kim, D. Bickhart, D. Null, T. Cooper, J. Cole, G. Wiggans, P. Ajmone-Marsan, L. Colli, E. Santus, G.E., Liu, S. Schroeder, L. Matukumalli, C. Van Tassell, and T. Sonstegard. 2013. Fine mapping for Weaver Syndrome in Brown Swiss cattle and the identification of 41 concordant mutations across NRCAM, PNPLA8 and CTTNBP2. PLoS One 8:e59251.

Rothammer, S., A. Capitan, E. Mullaart, D. Seichter, I. Russ, and I. Medugorac. 2014. The 80-kb DNA duplication on BTA1 is the only remaining candidate mutation for the polled phenotype of Friesian origin. Genet. Sel. Evol. 46:44.

Sargolzaei, M., Chesnais, J.P., and Schenkel, F.S. 2014. A new approach for efficient genotype imputation using information from relatives. BMC Genomics 15: 478.

Segelke, D., Chen, J., Liu, Z., Reinhardt, F., Thaller, G., Reents, R. 2012. Reliability of genomic prediction for German Holsteins using imputed genotypes from low-density chips. J. Dairy Sci. 95: 5403–5411.

Segelke, D., Täubert, H., Reinhardt, F., and Thaller, G. 2013. Chancen und Grenzen der Hornloszucht für die Rasse Deutsche Holstein. Zuchtungskunde 85: 4.

VanRaden, P.M., Null, D.J., Sargolzaei, M., Wiggans, G.R., Tooker, M.E., Cole, J.B., Sonstegard, T.S., Connor, E.E., Winters, M., van Kaam, J.B. 2013. Genomic imputation and evaluation using high density Holstein genotypes. J. Dairy Sci 96: 668–678.