Prediction of postpartum diseases of dairy cattle using machine learning

A.M. Hidalgo[1], F. Zouari[1], H. Knijn[1] & S. van der Beek[1]

[1] CRV, Wassenaarweg 20, 6843 NW, Arnhem, The Netherlands
Andre.hidalgo@crv4all.com (Corresponding Author)

Introduction

The postpartum period in dairy cows is the period in which most of the metabolic diseases occur (Goff & Horst, 1997; Ingvartsen, 2006). Diseases such as ketosis, milk fever and metritis arise in this time frame due to significant nutritional, metabolic and physiological changes that the cow undergoes. Proper management of this period in dairy cows is therefore essential to achieve optimal performance regarding production, reproduction and health; leading to an improved profitability (Mulligan & Doherty, 2007; von Keyserlingk *et al*., 2009).

Routinely large quantities of data is recorded in dairy cattle farms which can be used to improve management. Machine learning algorithms have been developed to learn from the data without the need of explicit modelling, therefore it is an excellent choice to handle such a big amount of information. Recently, machine learning algorithms have been applied in dairy cattle aiming at prediction of subclinical mastitis (Mammadova & Keskin, 2013) and insemination outcomes (Shahinfar *et al*., 2014; Hempstalk *et al*., 2015). Studies evaluating the use of machine learning to predict several postpartum diseases however are lacking. Therefore this study uses a machine learning algorithm to predict the probability of occurrence of postpartum diseases (60 days) in dairy cattle based on prepartum data (250 days).

Material and methods

Data

Data from 6 Dutch farms were used in this study; data were recorded from 2009 through 2016. These farms have well recorded health data. All animals used in the study had to have a calving date and a disease record; cows that did not have any health data registered during the entire period, were eliminated since we could not trust that information to produce the ground truth. The disease record as well as milk production record were combined with respect to a calving date. They had to be within a -250 and +60 days from a given calving date, so that it encompasses the postpartum period that we were interested as well as the previous period of the calving date from which information will be used to train the model. In total, 59,590 records from 1,470 animals were available. Forty six features were used to train the algorithm (Table 1). The disease status of an animal was defined as presenting at least one out of twenty diseases (metabolic, reproduction and infectious) that were recorded in the 60 days after calving. Data were not balanced for disease status, there were 47,428 non-sick (79.6%) and 11,862 sick (20.4%) animals. For features such as somatic cell count, parity and gestation length, bins were created based on biological information, whereas all other features were normalized.

Machine learning algorithm

Machine learning relies in the idea that computers find patterns and relationships in data without being manually programmed to do so. The computer will then develop a model that can be used for prediction. There are plenty of machine learning algorithms available, such as random forest, support vector machine (SVM), naïve Bayes, decision trees, neural networks, etc. In our study we used the random forest algorithm (Breiman, 2001) which is an ensemble learning method that uses weak classifiers to build a strong one. Briefly, this algorithm grows a "forest" out of a myriad of decision trees. Each tree is built from a random set of observations sampled with replacement from the training dataset. Each of these observations will also have a random set of features sampled. Prediction is then based on the most popular class voted by these decision trees. These votes are boosted and a probability is computed. The number of trees used to create a forest in this study was 200. Increasing the number of trees is always better with diminishing returns. We evaluated this threshold by changing the number of trees and monitoring the convergence of the AUC score with respect to the performance costs. The random forest algorithm was applied using the Apache Spark (Zaharia *et al*., 2016) library MLlib (Meng *et al*., 2016).

Table 1. Features used to train the algorithm.

Features	Count	Missing	Features	Count	Missing
Numerical			Breeding values		
Lactation length	59290	0	Overall index	59290	0
Exp.milk yield in 24h	47589	11701	Kg fat	59290	0
Milk yield	58181	1109	Kg protein	59290	0
Fat yield	58181	1109	Kg lactose	59290	0
Protein yield	58181	1109	Percentage fat	59290	0
Lactose yield	57946	1344	Percentage protein	59290	0
Fat yield 305 days	58181	1109	Percentage lactose	59290	0
Protein yield 305 days	58181	1109	Overall production	59290	0
Lactose yield 305 days	58034	1256	Calving ease	59290	0
Lactation index	56446	2844	Gestation length	59290	0
Number of high SCC	59290	0	Direct birth weight	59290	0
Milk yield 24h	47019	12271	Maternal calving ease	59290	0
Fat yield 24h	46735	12555	Maternal gestation length	59290	0
Protein yield 24h	46735	12555	Maternal birth weight	59290	0
Lactose yield 24h	46576	12714	Overall fertility	59290	0
SCC 24 hours	46804	12486	Non-return 56 days	59290	0
Urea concentration 24h	46475	12815	Interval calving-1st insem.	59290	0
Parity	59290	0	Calving interval	59290	0
Days in milk	58522	768	Interval 1st-last insem.	59290	0
			Conception rate	59290	0
Categorical					
Farm	59290	0	Sire ID	59290	11701
Color	59290	0	Dam ID	59290	0

Date of milk yield 24h	47589	0	Sickness status	59290	0
Diagnosis code	59290	0			

Validation

The complete dataset of 59,290 records was split into training (70%) and validation (30%) datasets. The training dataset contained 41,503 records from which 33,316 were not-sick and 8,187 were sick. The validation dataset consisted of 17,787 from which 14,112 were not-sick and 3,675 were sick. Down sampling was applied in both datasets so that the two categories contained balanced class distribution. The training dataset was then used to train the algorithm whereas we used the validation dataset to determine the predictive ability of the algorithm. The predictive ability was determined by computing the AUC score. A brief explanation of why it is oftentimes used for model assessment is described in Shahinfar *et al*. (2014). To test whether breeding values improve the prediction, we ran the analyses with and without them.

Results and discussion

Machine learning methods have recently started to be used for prediction in livestock. We obtained areas under the ROC curve of 0.77 and 0.81 (Figure 1) which shows that we could predict relatively well whether an animal will have a postpartum disease based on prepartum information (250 days).

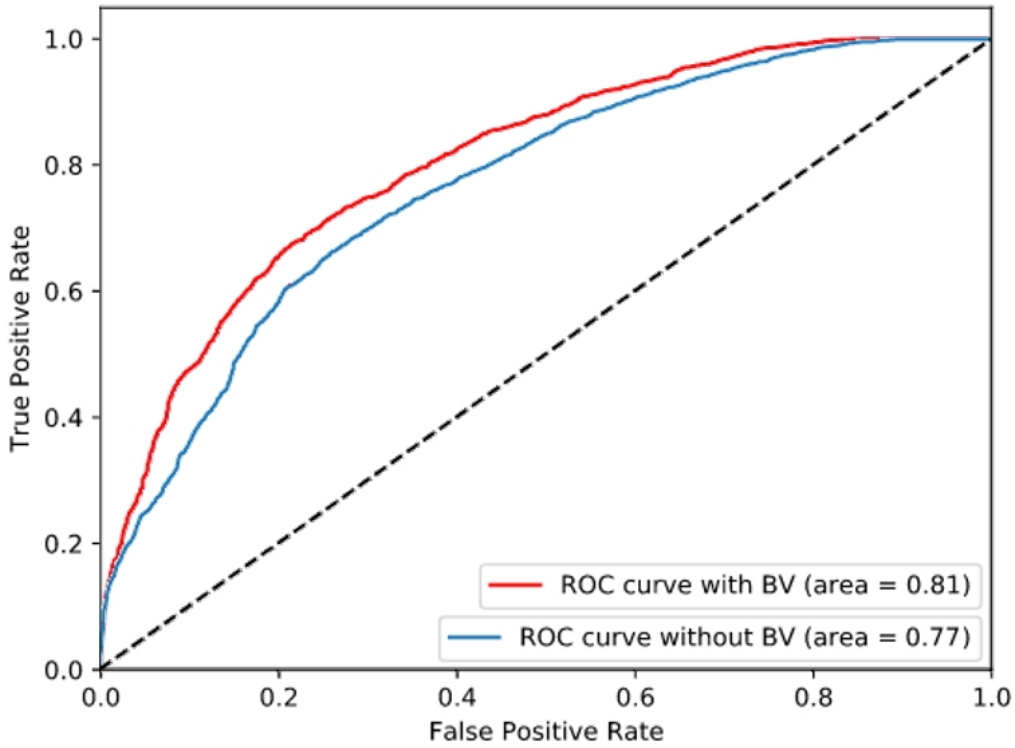

Figure 1. Area under the ROC curve (AUC) with and without the inclusion of breeding values.

The AUC score shows that adding breeding values to the model increases the predictive ability (Figure 1). The breeding values aided the trees in class separation (sick/not

sick). In the purpose of reducing overfitting, animal identification features were removed. This will allow the algorithm to examine the records without the risk of uncovering an underlying pattern in the animal identification numbers as a predictor. The end result of the prediction algorithm is however animal based, this is to say that we will be predicting an outcome that is suited for a particular animal. Breeding values help bridge this gap, by offering a digital representation of an animal with valid statistical value that can be used by the decision trees to better predict the outcome.

Recently, other studies have also performed prediction using machine learning in dairy cattle (Mammadova & Keskin, 2013; Shahinfar *et al*., 2014; Hempstalk *et al*., 2015). Mammadova & Keskin (2013) using lactation rank, milk yield, electrical conductivity, average milking duration, and control season were able to predict whether a cow had subclinical mastitis with an sensitivity of 89%, specificity of 92% and a surprisingly high error of 50% using SVM. However as they did not compute the AUC score, comparisons between results are difficult. Shahinfar *et al*. (2014) predicted the insemination outcome based on production, reproduction, health, and genetic information; using various machine learning algorithms. The AUC score that they found varied from 0.60 to 0.76 across all methods, from which naïve Bayes was the worst method and random forest was the best. Hempstalk *et al*. (2015) working on the prediction of conception success also studied several machine learning algorithms. Features in their model included production, reproduction, health, genetic information, and milk mid-infrared (MIR) data. They determined AUC score using only MIR only data, no MIR data, and using both data. The best results were achieved using no MIR data and varied from 0.613 to 0.675 and random forest was the best method. Comparing AUC scores from our study with other studies shows that we achieved a higher predictive ability than what currently is achieved by others, even though the target of prediction is different among studies.

Probability of postpartum diseases was predicted based on prepartum information fed to random forest algorithm with relatively good certainty. Such information can assist farmers to select which animals have to be more carefully evaluated and tested for postpartum diseases. We envisage that big data is out there to stay and it has different applications. We have tested one of them and it is very promising. It is also an important step to show that big data can and shall be used by breeding companies. Further on, other applications will be studied and developed always aiming at solutions that the farmer needs.

List of References

Breiman, L., 2001. Random Forests. Mach. Learn. 45:5-32.

Goff, J.P. & R.L. Horst, 1997. Physiological changes at parturition and their relationship to metabolic disorders. J. Dairy Sci. 80:1260-1268.

Hempstalk, K., S. McParland & D.P. Berry, 2015. Machine learning algorithms for the prediction of conception success to a given insemination in lactating dairy cows. J. Dairy Sci. 98:1-12.

Ingvartsen, K.L., 2006. Feeding- and management-related diseases in the transition cow: Physiological adaptations around calving and strategies to reduce feeding-related diseases. Anim. Feed Sci. Tech. 126:175-213.

von Keyserlingk, M.A.G., J. Rushen, A.M. de Passillé & D.M. Weary, 2009. Invited review: The welfare of dairy cattle – Key concepts and the role of science. J. Dairy Sci. 92:4101-4111.

Meng, X., J. Bradley, B. Yavuz, E. Sparks, S. Venkataraman, D. Liu, J. Freeman, D.B. Tsai,

M. Made, S. Owen, D. Xin, R. Xin, M.J. Franklin, R. Zadeh, M. Zaharia & A. Talwalkar, 2016. MLlib: Machine learning in Apache Spark. J. Mach. Learn. Res. 17:1-7.

Mammadova, N. & I. Keskin, 2013. Application of the support vector machine to predict subclinical mastitis in dairy cattle. Sci. World. J. 603897.

Mulligan, F.J. & M.L. Doherty, 2008. Production diseases of the transition cow. Vet. J. 176:3-9.

Shahinfar, S., D. Page, J. Guenther, V. Cabrera, P. Fricke & K. Weigel, 2014. Prediction of insemination outcomes in Holstein dairy cattle using alternative machine learning algorithms. J. Dairy Sci. 97:731-742.

Zaharia, M, R.S. Xin, P. Wendell, T. Das, M. Armbrust, A. Dave, X. Meng, J. Rosen, S. Venkataraman, M.J. Frankling, A. Ghodsi, J. Gonzalez, S. Shenker & I. Stoica, 2016. Apache Spark: A unified engine for big data processing. Commun. ACM. 59(11):56-65.