J. Dairy Sci. 97:731–742
http://dx.doi.org/10.3168/jds.2013-6693

Prediction of insemination outcomes in Holstein dairy cattle using alternative machine learning algorithms

Saleh Shahinfar,*[1] David Page,† Jerry Guenther,* Victor Cabrera,* Paul Fricke,* and Kent Weigel*
*Department of Dairy Science, and
†Department of Biostatistics and Medical Informatics and Department of Computer Science, University of Wisconsin, Madison 53706

ABSTRACT

When making the decision about whether or not to breed a given cow, knowledge about the expected outcome would have an economic impact on profitability of the breeding program and net income of the farm. The outcome of each breeding can be affected by many management and physiological features that vary between farms and interact with each other. Hence, the ability of machine learning algorithms to accommodate complex relationships in the data and missing values for explanatory variables makes these algorithms well suited for investigation of reproduction performance in dairy cattle. The objective of this study was to develop a user-friendly and intuitive on-farm tool to help farmers make reproduction management decisions. Several different machine learning algorithms were applied to predict the insemination outcomes of individual cows based on phenotypic and genotypic data. Data from 26 dairy farms in the Alta Genetics (Watertown, WI) Advantage Progeny Testing Program were used, representing a 10-yr period from 2000 to 2010. Health, reproduction, and production data were extracted from on-farm dairy management software, and estimated breeding values were downloaded from the US Department of Agriculture Agricultural Research Service Animal Improvement Programs Laboratory (Beltsville, MD) database. The edited data set consisted of 129,245 breeding records from primiparous Holstein cows and 195,128 breeding records from multiparous Holstein cows. Each data point in the final data set included 23 and 25 explanatory variables and 1 binary outcome for of 0.756 ± 0.005 and 0.736 ± 0.005 for primiparous and multiparous cows, respectively. The naïve Bayes algorithm, Bayesian network, and decision tree algorithms showed somewhat poorer classification performance. An information-based variable selection procedure identified herd average conception rate, incidence of ketosis, number of previous (failed) inseminations, days in milk at breeding, and mastitis as the most effective explanatory variables in predicting pregnancy outcome.
Key words: machine learning, reproductive management, dairy cattle

INTRODUCTION

Although it is often stated that the decline in reproductive performance of dairy cattle is due to intensive selection for high milk production, it is clear that many environmental features and management practices contribute directly to the insemination outcome. Management features, such as heat detection, nutrition, transition cow management, BCS, semen handling, metabolic disorders, udder health, calving difficulty, reproductive disease, and cow comfort strongly affect reproductive performance (Lucy, 2001; Caraviello et al., 2006; Schefers et al., 2010). Researchers have also reported associations between reproduction traits and genetics (Weigel, 2004; González-Recio and Alenda, 2005; Liu et al., 2008), milk yield (Berry et al., 2003; Windig et al., 2005, 2006; Tiezzi et al., 2011), heat stress (Morton et al., 2007), energy balance (de Vries and Veerkamp, 2000), timing of AI (Cornwell et al., 2006), reproductive health (Sheldon et al., 2002), lameness (Garbarino et al., 2004), quality and quantity of semen (Jaskowski and Szenfeld, 1999), sperm dosage in sex-sorted semen (DeJarnette et al., 2011), rump angle and conformation traits (Wall et al., 2005), and cow health (Chebel et al., 2004). Caraviello et al. (2006) used an alternating decision tree algorithm to identify frequency of hoof trimming, type of bedding in the dry cow pen, type of restraint system, and duration of the voluntary waiting period as key features in predicting first-service conception rate. They also found that bunk space per cow, temperature for thawing semen, percentage of cows with low BCS, number of cows in the maternity pen, strategy for using cleanup bulls, and milk yield at first service were the most informative variables in predicting the insemination outcome at 150 DIM.

Schefers et al., (2010) modeled conception rate and service rate of commercial dairy herds using a model tree algorithm. Their study identified percentage of repeated inseminations between 4 and 17 d post-AI

Received February 13, 2013.
Accepted September 11, 2013.
[1]Corresponding author: shahinfar@wisc.edu

(a measure of breeding protocol compliance), stocking density in the breeding pen, length of the voluntary waiting period, days from insemination to pregnancy check, and SCS as the most important features in predicting herd average conception rate. The most important explanatory variables for predicting herd average service rate were number of cows per breeding technician, resynchronization protocol, use of soakers in the holding area, and bunk space per cow in the breeding pens. The effects of negative energy balance in early lactation have been well studied and seem to be partially responsible for lower conception rates observed in high-producing cows. Oikonomou et al., (2008) showed that BCS, energy content of the diet, cumulative effective energy balance, and blood glucose have favorable genetic relationships with reproduction, whereas BHBA and NEFA are negatively correlated with energy balance and have unfavorable genetic correlations with reproductive traits. In that study, mean daily energy balance, milk protein content, and DMI during the first 28 d postpartum were associated with higher conception rate at first service, whereas cows with high DMI and positive energy balance had a shorter calving-to-conception interval. On the other hand, lower BCS have been associated with a longer calving-to-conception interval (Patton et al., 2007).

Although several studies have attempted to identify specific factors affecting insemination outcome in lactating dairy cattle, few have tried to predict the outcome of individual insemination events based on all health, reproduction, and production data available for each cow at the time of service. Obviously, such a prediction tool could be useful as a decision support system for dairy farmers.

The ability to accommodate large and complex data sets with missing values, as well as the lack of restrictive parametric assumptions, make machine learning methods good candidates for data mining and development of predictive tools in fields such as agriculture. Grzesiak et al. (2010) used artificial neural networks, multivariate adaptive regression splines, logistic regression, classification trees, and classification functions to classify cows with good or poor reproductive performance based on age, calving interval, gestation length, BCS, FCM, and average of fat and protein percentages. They reported classification accuracies of 85 to 86%, with sensitivity and specificity of 85%, for a multilayer perceptron with 2 hidden layers. Among the machine learning methods used in the animal sciences, artificial neural networks are the most frequently used, with applications such as predicting milk yield in dairy cows (Lacroix et al., 1995; Grzesiak et al., 2006; Gianola et al., 2011), classifying mastitis cases (Yang et al., 1999), classifying lameness in horses (Suchorski-Tremblay et al., 2001), predicting the slaughter weight of bull calves (Adamczyk et al., 2005), identifying SNP associated with chicken mortality (Long et al., 2009), and real-time prediction of breeding values in dairy cattle (Shahinfar et al., 2012).

The objective of this study was to compare the performance of different machine learning algorithms for predicting the insemination outcomes of lactating dairy cows using production, reproduction, health, and genetic information. Identification of specific environmental factors or management practices that affect reproductive performance is a by-product of the aforementioned analyses, but in this study our primary goal was to maximize predictive ability for development of a decision support tool.

MATERIALS AND METHODS

Data

The data used in this study were provided by 26 Wisconsin dairy farms that were enrolled in the Alta Genetics (Watertown, WI) Advantage Progeny Testing Program. A general description of these dairy herds can found in Table 1 of Schefers et al. (2010). After editing, the data set contained 129,245 breeding records from primiparous Holstein cows and 195,128 breeding records from multiparous Holstein cows. For each breeding event, there existed corresponding production data, EBV, health events, and reproduction information (Table 1). In terms of reproduction performance, herds in this study are representative of large commercial dairy farms in Wisconsin (Figure 1).

Production, reproduction, and health event data were obtained from backup files of the on-farm DairyCOMP 305 herd management software (Valley Agricultural Software, Tulare, CA) of individual farms. Estimated breeding values and calving ease data were extracted from the US Department of Agriculture Agricultural Research Service Animal Improvement Programs Laboratory (Beltsville, MD) database. Each data point in the final data set included 23 or 25 features and 1 binary response variable for primiparous or multiparous cows, respectively. Records were collected over a 10-yr period from 2000 to 2010.

To account for energy balance and reduce the dimensionality of features in the model for analysis, ECM was used as an explanatory variable. The following equation was used to determine the amount of energy needed for producing milk, adjusted to 3.5% fat and 3.2% true protein (Tyrrell and Reid, 1965):

$$\text{ECM} = [0.327 \times \text{milk (kg)}] + [12.95 \times \text{fat (kg)}] + [7.65 \times \text{protein (kg)}].$$