J. Dairy Sci. 100:1–12
https://doi.org/10.3168/jds.2016-11896

Novel approaches to assess the quality of fertility data stored in dairy herd management software

K. Hermans,*[1] W. Waegeman,† G. Opsomer,* B. Van Ranst,* J. De Koster,* M. Van Eetvelde,* and M. Hostens*
*Faculty of Veterinary Medicine, Department of Reproduction, Obstetrics and Herd Health, Ghent University, Salisburylaan 133, 9820 Merelbeke, Belgium
†Faculty of Bioscience Engineering, Department of Mathematical Modelling, Statistics and Bioinformatics, Ghent University, Coupure Links 653, 9000 Ghent, Belgium

ABSTRACT

Scientific journals and popular press magazines are littered with articles in which the authors use data from dairy herd management software. Almost none of such papers include data cleaning and data quality assessment in their study design despite this being a very critical step during data mining. This paper presents 2 novel data cleaning methods that permit identification of animals with good and bad data quality. The first method is a deterministic or rule-based data cleaning method. Reproduction and mutation or life-changing events such as birth and death were converted to a symbolic (alphabetical letter) representation and split into triplets (3-letter code). The triplets were manually labeled as physiologically correct, suspicious, or impossible. The deterministic data cleaning method was applied to assess the quality of data stored in dairy herd management from 26 farms enrolled in the herd health management program from the Faculty of Veterinary Medicine Ghent University, Belgium. In total, 150,443 triplets were created, 65.4% were labeled as correct, 17.4% as suspicious, and 17.2% as impossible. The second method, a probabilistic method, uses a machine learning algorithm (random forests) to predict the correctness of fertility and mutation events in an early stage of data cleaning. The prediction accuracy of the random forests algorithm was compared with a classical linear statistical method (penalized logistic regression), outperforming the latter substantially, with a superior receiver operating characteristic curve and a higher accuracy (89 vs. 72%). From those results, we conclude that the triplet method can be used to assess the quality of reproduction data stored in dairy herd management software and that a machine learning technique such as random forests is capable of predicting the correctness of fertility data.
Key words: dairy reproduction, data quality, dairy herd management software, random forests

INTRODUCTION

Researchers often use data stored in dairy herd management software to facilitate the collection of fertility data (Caraviello et al., 2006; Zwald et al., 2006; Shahinfar et al., 2014). These are so-called secondary data sources, referring to the fact that the data are collected by someone other than the user (i.e., the researcher) and not specifically collected for research purposes. Conclusions based on data from dairy herd management software are regularly published in scientific journals (Zwald et al., 2004; Caraviello et al., 2006; Wenz and Giebel, 2012) without proper data quality assessment or data cleaning (Harpe, 2009). Despite the increasing importance of data quality (Wang and Strong, 1996; Arts et al., 2002) and the rich theoretical and practical contributions in all aspects of data cleaning (Ballou and Pazer, 1985; Wand and Wang, 1996; Pipino et al., 2002), no single end-to-end off-the-shelf solution is available to automate the detection of incorrect data. Often a significant portion of the cleaning work has to be done manually or by low-level programs (Rahm and Do, 2000), making data quality assessment an expensive and time-consuming process (Wang et al., 1995; Haug et al., 2011). In addition, the ability to gather data via handheld computers, as well as more complex data capturing systems working in tandem with technologies such as voluntary milking systems and heat detection aids, has outpaced the speed and cost of convenient data quality assessment.

The first objective of the present study was to develop a deterministic or rule-based data cleaning method that is easy to understand and quick to implement. A novel method for screening physiologically plausible or implausible sequences of reproduction events was introduced. The time series events for every animal

Received August 19, 2016.
Accepted January 4, 2017.
[1]Corresponding author: kristofhermans@bovinet.be

were converted to a symbolic representation and split into triplets (3-letter code). To this end, triplets were manually labeled as physiological correct, suspicious, or impossible. As such it becomes possible to determine whether or not data from an individual cow are complete and correct to serve as input for statistical analysis.

The second objective of the present study was to develop a probabilistic or prediction-based data cleaning method based on cow- and farm-related variables. The probabilistic data cleaning method could be used to preselect cow records with a high probability of being correct at the time of data extraction and in the future (early phase data assessment). Hence, time and cost needed for deterministic data cleaning could decline (Haug et al., 2011).

To do so, we adopted a machine learning methodology, random forests (**RF**; Breiman, 2001), by searching for statistical relationships between data correctness and certain predictor variables that are characteristic for an individual cow. To this end, we first defined what we consider as correct and incorrect data records. This definition was used to manually label the records from dairy cows on 26 farms. The database was split into a training set, which served as input to train the RF, and a testing data set, which served to validate the prediction performance of the RF.

In our experimental results, the detection performance of RF (Breiman, 2001) is compared with a classical linear statistical method (penalized logistic regression, **PLR**; Loeffler et al., 1999; Fourichon et al., 2000; López-Gatius et al., 2005). Machine learning algorithms are able to accommodate for complex nonlinear relationships within data; for that reason, their prediction capabilities often outperform classical statistical methods (Lim et al., 2000; Loh, 2011). Additionally, variables affecting the quality of reproductive data from dairy cows are identified. These variables could be of interest for including or excluding farms and cows for government, disease control, or research purposes.

MATERIALS AND METHODS

Data Extraction

Twenty-six farms enrolled in the herd health management program from the Faculty of Veterinary Medicine, Ghent University, Belgium, and those possessing a dairy management software program, were included in the study. Backup files from the herd management software used on these farms were extracted by the Dairy Data Warehouse (2016; **DDW**) and stored without manipulation of the data. Reproduction and mutation or life-changing events like birth, purchase, sale, or death data were obtained from DDW via a web application programming interface. Reproduction events encountered in the present study were heat, insemination, positive pregnancy check, negative pregnancy check, do not breed, and abortion. Natural mating, AI, and embryo transfer were all classified as an insemination event. No other data than mutation and fertility events were included in the study.

The data provided via the DDW application programming interface were loaded into the statistical program R version 3.2.5 (Ripley, 2001) and merged into one file. Male animals were excluded from the data set. Events were grouped by farm and cow identification number before being chronologically sorted. Events from the same cow that were recorded on different farms (due to selling and buying) were not merged.

Deterministic Data Cleaning Method

Triplet Creation. An alphabetical letter code (Table 1) was assigned to every fertility and mutation event to convert the time series data into a symbolic form (Aref et al., 2004). A single letter strand was configured for each cow by combining all letter codes (e.g., **BHICH** representing a birth, heat, insemination, calving, and heat event). The single letter strand was constructed in a chronological direction, from the oldest to the newest event (Figure 1). The letter strand was then divided into triplets. A triplet was defined as a letter sequence representing 3 consecutive events (Figure 1). By using 3 consecutive events instead of 2, the relationship between the first and third event could be checked as well. For example, many cows show heat signs during pregnancy; accordingly, a farmer can record a heat event after a positive pregnancy check (PH; Sturman et al., 2000, Roelofs et al., 2010). The recording of a do not breed event after a heat event is also plausible. However, the sequence of events turns out to be suspicious when the 3 events are combined (PHD). Adding an extra event to the triplet sequence, creating a quadruple, would complicate interpretation as the number of relational combinations would increase 10-fold.

The first triplet started at the oldest event and stopped when 2 consecutive events were added (e.g., BHI is the first triplet from the letter strand BHICH). The next triplet within the same letter strand starts by moving up one place (e.g., HIC is the second triplet from the letter strand BHICH). This process was repeated until the last event of the letter strand has been used in a triplet (e.g., ICH being the last triplet from the letter strand BHICH). An exception to the triplet definition was made if the total number of events was less than 3. If only 1 or 2 events were recorded, then