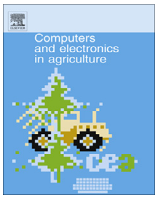
Computers and Electronics in Agriculture 127 (2016) 26–37

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

journal homepage: www.elsevier.com/locate/compag



Original papers

Recognition of culling reasons in Polish dairy cows using data mining methods

Krzysztof Adamczyk[a,*], Daniel Zaborski[b], Wilhelm Grzesiak[b], Joanna Makulska[a], Wojciech Jagusiak[c]

[a] University of Agriculture in Krakow, Institute of Animal Sciences, Department of Cattle Breeding, 30-059 Kraków, Al. Mickiewicza 24/28, Poland
[b] West Pomeranian University of Technology in Szczecin, Laboratory of Biostatistics, 71-466 Szczecin, Doktora Judyma 10, Poland
[c] University of Agriculture in Krakow, Department of Genetics and Animal Breeding, 30-059 Kraków, Al. Mickiewicza 24/28, Poland

ARTICLE INFO

Article history:
Received 18 September 2015
Received in revised form 19 May 2016
Accepted 25 May 2016

Keywords:
Dairy cows
Involuntary culling
Data mining methods

ABSTRACT

Cow longevity and reasons for culling are one of the most important research problems in the contemporary cattle breeding. Therefore, the analysis of the relationship between cow performance and involuntary disposal contributes to taking more informed decisions in herd management. The aim of the present study was to compare the efficiency of artificial neural networks (ANN) and boosted classification trees (BT) with that of linear discriminant analysis (LDA) and classification functions (CF) in recognizing culling reasons of dairy cows in Poland, based on the lifetime performance data, routinely monitored in a herd. The analyses carried out in the present study showed that the accurate recognition of different culling reasons based on predictors included in the above-mentioned models is, in general, impossible. Only BT had limited discrimination abilities, but the results obtained using this method were not much improved compared with ANN and LDA with CF. In order to predict precisely various culling reasons, more specific data are required. They could be obtained from the increasingly popular, technologically advanced, systems of real-time monitoring of animal health status (physical activity, rumination rate, etc.), dependent also on environmental conditions (e.g. temperature–humidity index).

1. Introduction

Long-term selection for production traits and intensification of dairy production systems resulted in a high milk performance. However, simultaneously, a noticeable deterioration of fertility and health traits, contributing to the decreased cow longevity and overall herd profitability, was observed. Profitability of dairy herd is determined mainly by milk yield per lactation and the length of cow productive life which in intensive production system often does not exceed three lactations. It makes that many cows do not produce at a mature level and their genetic potential is not fully utilized (Knaus, 2009). In 2014 the average length of productive life of milk recorded cows in Poland was 3.12 years and the average lifetime production amounted to 22,722 kg of milk (PFCBDF, 2015). Involuntary culling in high milking dairy herds significantly limits the optimum decisions on replacing cows with heifers of higher genetic potential.

Cow longevity and most reasons for involuntary culling are associated with production intensity which is mainly decided by breeders (Ahlman et al., 2012). The herd management decisions significantly affect not only the achieved profitability but sometimes even the future existence of the herd. Excessive intensification without providing adequate animal welfare may result in the decreased fertility and health problems which are very important reasons for involuntary disposal, directly or indirectly – by generating the additional costs that make further use of a cow as unprofitable (Groenendaal et al., 2004; Langford and Stott, 2012).

In order to improve cow longevity in the contemporary cattle breeding programs more balanced breeding goals, incorporating functional traits, are defined (Boichard and Brochard, 2012; Axelsson, 2013). At the same time a systematic improvement of livestock farming conditions to maintain an optimal level of animal welfare is observed (Nielsen et al., 2006; Forabosco et al., 2009; Oltenacu and Broom, 2010). The studies of Langford and Stott (2012) indicate that improving cow welfare by reducing mastitis, lameness or infertility over the long term positively affects the mean longevity in a dairy herd. Since the vast majority of cow disease events take place in the so-called transition period the proper herd management in this time is crucial for subsequent cow performance and survival. To assess the quality of transition management practices in American freestall herds Transition Cow Index™ was used by Nordlund (2013). The possibilities to estimate the

---

* Corresponding author.
E-mail address: rzadamcz@cyfronet.pl (K. Adamczyk).

http://dx.doi.org/10.1016/j.compag.2016.05.011
0168-1699/

level of dairy cow welfare based on routine herd data, collected in national databases, were discussed by de Vries et al. (2011).

The knowledge of the association of various indicators monitored at the individual animal and herd level with cow lifespan and the reasons of involuntary culling is essential to predict cow longevity and to support optimal decisions in herd management. Based on the hitherto research two groups of longevity indicators – metabolic indicators and performance traits, routinely recorded in dairy cattle, can be distinguished. For example, Seifi et al. (2011) and Roberts et al. (2012) proposed monitoring of serum non-esterified fatty acids, β-hydroxybutyrate and calcium concentrations in the cow perinatal period in order to detect early health problems (mainly metabolic) that could be the reason of involuntary culling in the first 60 days of lactation. The possibilities to estimate a genetic merit for productive life from milk production and linear type traits were described by Weigel et al. (1998) and Caraviello et al. (2004a). In the last years, the studies on the genetic conditioning of longevity were also carried out. Szyda et al. (2011) found the association between the polymorphisms within the leptin gene and longevity of Holstein–Friesian cows, so these polymorphisms can be considered as the potential markers in cattle selection.

To predict cow longevity and support economical culling decisions relatively numerous quantitative techniques have been used (Lehenbauer and Oltjen, 1998). Congleton (1988) developed a model to evaluate future income from milk production and financial risk associated with culling decision. Caraviello et al. (2004a, 2004b) proved usefulness of survival analysis to estimate breeding value for longevity in Holstein–Friesian and Jersey bulls based on 305d lactation of the cows related with them. Langford and Stott (2012) applied a dynamic programme model to explore the relationships between financial outcomes, investment in improving welfare, cow lifespan and involuntary culling in dairy systems.

While the papers on searching associations between performance traits and dairy cow longevity are relatively numerous, the number of publications dealing with the relationship between routinely monitored traits and reasons of involuntary culling are rather scarce (Lacroix et al., 1997, for example). What is more, these relationships can be either linear or non-linear, which makes that their analysis requires methods capable to capture and represent complex relationships. Such methods include, among others, artificial neural networks, increasingly used in livestock breeding and management (Gardner et al., 1999; Grzesiak et al., 2006; Pastell and Kujala, 2007; Edriss et al., 2008; Njubi et al., 2010; Gianola et al., 2011) as well as less popular but still powerful boosted classification trees (Zaborski et al., 2014).

In the present study the effectiveness of the artificial neural network (ANN) and boosted classification trees (BT) to recognize culling reasons of dairy cows in Poland, based on the variables routinely monitored in a herd, was compared with the effectiveness obtained using a more traditional statistical method, which is the linear discriminant analysis (LDA) with classification functions (CF).

2. Material and methods

2.1. Animals

A total of 136,514 culling records of dairy cows were used. All animals were culled in 2012. Different breeds in the original data set were subsequently assigned to two categories: HF (Holstein–Friesian Black-and-White and Holstein–Friesian Red-and-White) and Other (Jersey, Simmental, Montbeliard, Polish Black-and-White, Polish Red-and-White, Polish Red, crossbreds of dairy breeds and crossbreds of dairy and beef breeds).

2.2. Input and output data

The whole data set was divided into three subsets according to the herd size (HS1 – 2–50 cows, HS2 – 51–100 cows, HS3 – 101–1075 cows). The input consisted of ten continuous and one categorical independent variables, used for recognition of culling reasons. The continuous variables were: DIM – lifetime number of days in milk (days), MILK – lifetime milk production (kg), FAT – lifetime fat production (kg), PROT – lifetime protein production (kg), CALA – age at first calving (yr), CULA – culling age (yr), FATP – a mean lifetime fat percentage in milk (%), PROTP – a mean lifetime protein percentage in milk (%), MILKD – a mean milk yield per milking day (kg), ECMD – a mean energy corrected milk yield per milking day (kg). ECMD was computed according to a slightly modified formula of Sjaunja et al. (1990):

$$\mathrm{ECMD} = (\mathrm{MILK}/\mathrm{DIM}) \times (383 \times \mathrm{FATP} + 242 \times \mathrm{PROTP} + 783.2)/3140 \tag{1}$$

The categorical variable was breed of a culled cow (BREED).

The output (dependent variable) was culling reason. In the initial data set, ten categories of culling reason were distinguished (code numbers according to the national milk recording system): C7 – low milk yield, C8 – udder diseases, C9 – reproductive disorders, C10 – contagious diseases (including leukemia), C11 – old age, C12 – metabolic and digestive system diseases, C13 – respiratory system diseases, C14 – locomotor system diseases, C15 – accidents, C16 – others (Słoniewski, 2010). Subsequently, four categories (C7, C10, C11, C13) related to the least frequent causes of culling (not exceeding 5% of the total number of culling cases) were removed from the data set.

Means and standard deviations of continuous variables and distribution of breeds and culling reasons in each herd-size subset are given in Tables 1 and 2, respectively.

2.3. Statistical analyses

Four different types of artificial neural networks were used for classification, i.e. linear networks (LN), radial basis function (RBF) networks and multilayer perceptrons (MLP) with one (MLP1) and two (MLP2) hidden layers. In order to train ANN, each herd-size subset was randomly divided into a training (L), validation (V) and test (T) set (the sizes of the subsets are shown in Table 1). The first set was used for the proper training of ANN, the second for the current monitoring of the training process and prevention of over-training and the last one for the final assessment of the network performance on completely new data, not used previously during the model construction. Before being fed to the input layer of the network, the numeric inputs were rescaled to the proper interval [0, 1] with the min–max method, whereas the categorical input (BREED) was converted to a binary form. On the other hand, the outputs were encoded with the one-of-*N* method, i.e. the number of neurons in the output layer was equal to the number of culling categories in the output variable. The classifications in the output layer were carried out using the acceptance and rejection thresholds, with each case being assigned to the most probable class (the option of complete classification). The networks were trained until reaching the lowest root-mean-squared error (RMSE) on the V set, which was calculated according to the following formula:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - y_{ip})^2} \tag{2}$$

where n is the number of cases in a given set, y_i is a real value, and y_{ip} is a predicted value.