Computers and Electronics in Agriculture 99 (2013) 1–6

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

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

Review

Predicting mastitis in dairy cows using neural networks and generalized additive models: A comparison

S. Ankinakatte [a,*], E. Norberg [a], P. Løvendahl [a], D. Edwards [a], S. Højsgaard [b]

[a] Department of Molecular Biology and Genetics, Aarhus University, Blichers Allè 20, 8830 Tjele, Denmark
[b] Department of Mathematical Sciences, Aalborg University, 9220 Aalborg Ø, Denmark

ARTICLE INFO

Article history:
Received 12 March 2013
Received in revised form 24 June 2013
Accepted 26 August 2013

Keywords:
Mastitis
Online data
Modeling techniques

ABSTRACT

The aim of this paper is to develop and compare methods for early detection of oncoming mastitis with automated recorded data. The data were collected at the Danish Cattle Research Center (Tjele, Denmark). As indicators of mastitis, electrical conductivity (EC), somatic cell scores (SCS), lactate dehydrogenase (LDH), and milk yield are considered. Each indicator is decomposed into a long-term, smoothed component, and a short-term, residual component, in order to distinguish long-term trends from short-term departures from these trends. We also study whether it is useful to derive a latent variable that combines residual components into a score to improve the model. To develop and verify the model, the data are randomly divided into training and validation data sets. To predict the occurrence of mastitis, neural network models (NNs) and generalized additive models (GAMs) are developed using the training set. Their performance is evaluated on the validation data set in terms of sensitivity and specificity. Overall, the performance of NNs and GAMs is similar, with neither method appearing to be decisively superior. NNs appear to be marginally better for high specificities. NNs model results in better classification with all indicators, using individual residuals rather than factor scores. When SCS is excluded, GAMs shows better classification result when milk yield is also excluded. In conclusion, the study shows that NNs and GAMs are similar in their ability to detect mastitis, a sensitivity of almost 75% observed for 80% of fixed specificity. Including SCS in the models improves their predictive $\geqslant$5% ability.

Contents

* Corresponding author. Tel.: +45 8715 7931.
E-mail address: Smitha.AA@agrsci.dk (S. Ankinakatte).

0168-1699/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.compag.2013.08.024

1. Introduction

Mastitis is the most costly disease affecting dairy cattle production and has a negative impact on animal welfare and product quality (Halasa et al., 2007; Hogeveen et al., 2011). Therefore it is of great interest to detect an outbreak of the disease as early as possible in order to be able to start treatment and separate the milk. The study was concerned with the clinical mastitis (mastitis hereafter). The ongoing introduction of new technologies for monitoring dairy cattle provides online measurements that may be useful for this purpose.

Mastitis is associated with many changes in the cow and in milk, and a combination of more than one indicator has been proven to be useful (Hogeveen et al., 2010). Somatic cell score (SCS) is historically the most commonly used indicator of mastitis, but other indicators, such as electrical conductivity (EC) and lactate dehydrogenase (LDH), have been introduced in the last decades. Measuring the indicators like SCS, and LDH, involves laboratory expenses. Whereas, EC, and milk yield, involves no expenses after initial machinery setup.

Prediction of mastitis has been studied by several authors. Nielen et al. (1995) used a neural network (NNs) model to differentiate between healthy and mastitic cows based on EC measurements, and suggested that this could be improved using additional indicators. Norberg et al. (2004) studied the ability, based on various traits derived from EC measurements, to predict mastitis singly and in combination using discriminant analysis. In particular the interquarter ratio of EC was found to have good predictive value. Using a linear mixed model approach, Chagunda et al. (2006a) developed a measure of mastitis risk based on LDH, somatic cell counts (SCC), and additional factors such as, days from calving, breed, parity, milk yield, and udder characteristics. Højsgaard and Friggens (2010) developed a state-space model for the degree of infection using SCC, EC, and LDH, and also discussed the use of milk yield as an indicator of mastitis, as reduction in milk yield is associated with mastitis (Green et al., 2006). Rutten et al. (2013) has reviewed the importance of using sensor data in dairy management. Another study by Chagunda et al. (2006b) showed that days in milking (DIM) was a significant predictor ($p < 0.001$) of LDH activity and SCS.

As with previous studies, the aim of this paper is to build a dynamic model using a panel of indicators recorded in the current study, for the early detection of clinical mastitis. The approach adopted in the current study is novel in several respects. (1) Each indicator is decomposed into a smoothed component and a residual component, in order to distinguish long-term trends from short-term departures from these trends. (2) Residual components are used to construct a latent variable (or score) to predict mastitis. (3) Finally, classifiers are constructed and compared based on two methodologies: NNs (Ripley, 1996) and GAMs (Hastie and Tibshirani, 1990).

This study compares the two modeling approaches for predicting the occurrence of mastitis, and compares the use of different indicators of mastitis in the model. In general no single indicator captures all aspects of mastitis (Hammann, 2005), thus panels of indicators of mastitis are considered. Milk yield has been introduced in the model to see whether it contributes to the detection of mastitis. In the panel of indicators, SCS is the most expensive indicator to measure. Of particular interest was whether SCS can be omitted from the panel of indicators without substantial loss of predictive ability.

2. Materials and methods

2.1. Data

The data used for development and validation of the model were collected at the Danish Cattle Research Centre (Tjele, Denmark), between October 2003 and November 2006. All cows were milked with an automatic milking system (AMS). The data consisted of a multivariate, irregularly spaced time series data from 401 cows with 664 lactations. There were 113 cows which have been treated for mastitis at least once. In total 346 mastitis treatments were recorded in 141 lactations. A large proportion of the lactations were uninformative and needed to be filtered from the data for further analysis. In the initial stage, the lactations with minimum of 200 milkings and in which all indicators were recorded at least 10% of total milking were considered. This filtering gave 87 of 141 treated lactations, which were considered to develop the classifiers.

The decision to give mastitis treatment to a cow was based on the California Mastitis Test (CMT) (Carroll and Schalm, 1962), which in turn was based on measurements on the monitoring tool. The monitoring tool measures the performance (milk yield) and electrical conductivity. An alarm was raised if there was a discrepancy in EC or milk yield of more than 15% from 4 quarters average. If an alarm was raised, the cow remained on the attention list for a small number of days, depending on the severity of the alarm. If the discrepancy from the 4 quarters average decreased, then the cow's udder health status as considered to be improving. Otherwise the CMT was performed, in order to make a firm decision before giving treatment for clinical mastitis (thus subclinical mastitis were never treated). The CMT provides a qualitative estimate of the SCS in the foremilk of the individual cows or the quarters.

For the estimation of population specific parameters, two subsets of the data were identified; the control set and the treated set. The control set consisted of lactations in which the cows were not treated for mastitis and the treated set consisted of lactations in which the cows were treated for mastitis. Some parameters were estimated using data from the control set, subject to the condition that in the selected lactations the cows must have had more than 800 milkings. Ten lactations, consisting 10,712 milking records, were selected for estimation of parameters. The estimated parameters were used for the prediction in treated set.

The data from the treated lactations were considered to develop the classification methods. The lactations to develop the classification methods were selected with some conditions: lactations should have had at least 300 milkings, and the proportion of LDH recorded per lactation should be more than 20% of the total milkings. This gave a subset of 69 lactations with 58,422 milkings. The subset was randomly divided into equal proportions, in terms of number of milkings, into a training set and a validation set.

The state of mastitis was considered 0 or 1, 0 for being healthy and 1 for being infected. The health definitions were made based on the day of mastitis treatment to the cow. The cow was considered to be in the state of mastitis in the period 3 days before and 3 days after the day of treatment. If the cow was treated in the successive day of the previous treatment, then it was considered as single treatment.

2.1.1. Data editing

Prior to the analysis some pre-processing was performed as follows. The SCS was the logarithm of somatic cell counts (cells/mL). The LDH was the activity of LDH in milk (μmol/min/liter). The EC was the interquarter ratio of EC (mS) as calculated by Norberg et al. (2004). The total milk yield per milking from all the quarters was transformed to yield per 24 h as explained in Section 2.2.1. The milkings with milking frequency, the number of milkings per 24 h, less than 1 were considered outliers and removed from the data.