Comparison of data driven mastitis detection methods

D. Jensen[1], M. van der Voort[2], C. Kamphuis[3], I.N. Athanasiadis[4], A. De Vries[5] and H. Hogeveen[2]
[1]*Department of Veterinary and Animal Sciences, University of Copenhagen, 1870 Frederiksberg C, Denmark*
[2]*Business Economics Group, Wageningen University& Research, 6706 KN Wageningen , The Netherlands*
[3]*Animal Breeding & Genomics, Wageningen University & Research, 6708 PB Wageningen , The Netherlands*
[4]*Information Technology Group, Wageningen University & Research, 6706 KN Wageningen, The Netherlands*
[5] *Department of Animal Sciences, University of Florida, Gainesville 32611-0910, United States*
Mariska.vandervoort@wur.nl

Abstract

The aim of this study is to compare the performances of different data driven methods for their ability in early detection of clinical mastitis. Many scientific papers on data driven methods for early mastitis detection have been published in the last decade. The performances vary greatly as well as the data use, time window, and gold standard definition. To be able to compare the performances of these data driven methods, this study applied various data driven methods including time series filtering and classification methods (i.e. Naïve Bayesian networks and Random Forest) under similar conditions. Forecast errors and filtered means of the time series models were used to distinguish mastitis cases from non-cases. Moreover, we focused solely on electrical conductivity (EC) measures of milk to detect clinical mastitis. Data for this study were provided by Lely Industries and originate from 57 farms in six different European countries with a total of 1,094,780 cow milkings with EC measurements at quarter milk level. It is hypothesised that the performances with respect to mastitis detection will differ substantially between the different methods, and that the ranking of methods is not consistent across different datasets. Despite, our preliminary results suggest that the performances of Naïve Bayesian networks and Random Forest do not vary much. The various filtering methods also present similar results. Although our naive approach of data handling allows us to compare different methods, we expect that each method in itself will improve when other variables than EC are included.

Keywords: mastitis, classification, transformation, filtering, EC

Introduction

Sensors generate a large amount of data, but as such do not provide any information on which decisions can or should be taken. With the development of mastitis sensor systems, an increasing number of scientific papers on early mastitis detection are being published (Hogeveen *et al.*, 2010), and various data driven methods are applied to translate sensor data into useful information for mastitis detection (Dominiak and Kristensen, 2017). Not only do these publications report a wide range of applied methods, but they also use a

wide variety of gold standards for mastitis, time windows for detection, and the selection of sensor data. This makes the studies difficult, if not impossible, to compare.

Timely detection of mastitis is of interest from an animal health and welfare perspective, but also plays an important economic role. Milk electrical conductivity (EC) is the most commonly used sensor data to detect clinical mastitis (De Mol, and Ouweltjes, 2001; Khatun *et al.,* 2017). But also other sensor data like milk colour, somatic cell count, and milk yield have been used to classify abnormal milk, often caused by clinical mastitis (Ebrahimie *et al.*, 2018). For the early detection of mastitis, alerts need to be generated. This is commonly achieved by applying methods that can produce an alert when the measured sensor data (that are considered a proxy for cow health status) deviates from the expected measurements (being a proxy for a normal, i.e. healthy status).

In the search for a perfect alert (that is, all mastitis cases receive an alert in time with 100% positive predictive value) a combination of different data driven methods have been used. Roughly we can distinguish three method families: filter methods, transformation methods, and classification methods. Filtering is a pre-processing method that defines, detect and correct errors of raw sensor data to minimize the impact of these errors on the succeeding subsequent analyses. Filtering methods are used to, remove noise from time series measurements and thus highlight the underlying trends from a signal and estimate the true underlying value. Transformation methods are also referred as a pre-processing step which makes input data more amendable by changing a range of number from one representation to another. Filtering or transformation methods to the data can result in more suitable parameters to be used for classification. Classification methods are used to convert the sensor data into and alert for mastitis detection.

The accuracy of the classification methods are evaluated by calculating the specificity (Sp) and sensitivity (Se), which are statistical measures of a binary classification test. Studies report various performance levels, ranging between 69-99% for Sp and 32-100% for Se (Hogeveen *et al.*, 2010, Dominiak and Kristensen, 2017). Most of these studies used a combination of sensor and non-sensor data. When based solely on sensor data, the detection of all clinical mastitis cases, with a manageable number of false positive attentions, is challenging.

Improving the performances of classification methods can be achieved by combining different data sources and changing the time-window. The objective of this paper is to evaluate the performances of several data driven methods for the early detection of clinical mastitis under similar conditions (i.e. data input, data selection criteria, time-window and gold standard). With this paper we want to strengthen the knowledge on the performance of different methods in relation to mastitis detection.

Material and methods

Data management

For this study we used sensor data on EC and somatic cell count (SCC), which had been automatically recorded using Lely milking robots, along with information to identify the individual cow and the herd it came from. Lely Industries (Maassluis, the Netherlands) provided these data.

We included a total of 296,501 records from 344 individual cows, encompassing 57 farms in four different European countries. Each individual cow was only represented with a single lactation. The included records were all made on days in milk (DIM) between 4 and 305. Only cows where SCC had been recorded at least once per week during this period were included in this study.
A SCC above 150,000 cells/mL was considered elevated for primiparous cows, and 250,000 cells/mL was considered elevated for multiparous cows, in accordance with Dutch standard practice based on the paper by Schepers et al. (1997). A cow was defined as having a mastitis event (ME) based on Kamphuis et al. (2016): at a given observation time if at least two of the three most recent milkings showed elevated SCC. A single mastitis event was not limited to three consecutive milkings, but could continue as long SCC is elevated. The ME start at the milking where SCC is elevated for the first time, and ends at the last observation of elevated SCC, followed by four observation without SCC elevation.

From the 344 cows, 19 cows did not experience any ME. The cows, which did experience mastitis at least once (N = 325), were divided into a training set for training different classification methods, and a test set for testing the models. This division was based on farm, where 2/3 of the farms were randomly selected to be used in the training set, and the remaining 1/3 were used as the test set. This division by farm was done to ensure independence between training and test data.

Time series filtering

For this study, we implemented a total of four different time series filtering methods, which are commonly used in the scientific literature relating to precision livestock farming. These filtering methods were optimized on the 19 cows without any ME in their lactation. The filtering methods were, in order of increasing complexity: 1) a moving average (MA), 2) an exponentially weighted moving average (EWMA), 3) a univariate dynamic linear model (DLM), and 4) a multivariate DLM. These were all implemented in R (R Core Team, 2017). Each of the filtering methods was optimized by finding the value of the relevant variables (see below), which minimized the root of the mean squared errors (RMSE) when applied to the filtering optimization data.

Moving Average (MA)

At each time step, the filtered value, z_t, is defined as the simple mean of the n most recently observed values, with n being a predefined integer value called the "window length". The forecast for the observation at a given time t is given as the filtered value, z_{t-1}, at time t-1. The forecast variance is estimated as

$$\sigma_{z_t}^2 \approx \frac{\sigma^2}{n} \quad (1)$$

, where σ^2 is the variance of the observed values and n is the window length. The MA was optimized for this study by trying values of n between 1 and 10 by steps of 1.

Exponentially weighted moving average (EWMA)

At each time step, the filtered value is defined according to the following equation:

$$z_t = \lambda \cdot k_t + (1 - \lambda) \cdot z_{t-1} \quad (2)$$

, where λ is a scale factor which can take values between 0 and 1, and k_t is the observed value at time *t*. The forecast for the observation at a given time *t* is given as the filtered value, z_{t-1}, at time *t*-1. The forecast variance is estimated as

$$\sigma_{z_t}^2 \approx \sigma^2 \cdot \left(\frac{\lambda}{2 - \lambda}\right) \quad (3)$$

, where σ^2 is the variance of the observed values. The EWMA was optimized by trying values of λ between 0 and 1 by steps of 0.01.

The univariate and multivariate dynamic linear model (DLM)

For this study, we implemented first-order univariate and multivariate DLMs without systematic growth components. At each time step, the EC values are filtered using the Kalman filter, as described in detail by West & Harrison (1997). The filtered values of one (univariate) or four (multivariate) EC values at time step t are defined by the system equation:

$$\boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1} + w_t \quad (4)$$

, where $\boldsymbol{\theta}_t$ is the parameter vector, and the error term is defined as $w_t \approx N(\underline{0}, \boldsymbol{W})$ with $\boldsymbol{W}$ being the systematic co-variance matrix. In our implementation, $\boldsymbol{W}$ was estimated continuously during the Kalman filtering by means of a discount factor, δ, which can take values between 0 and 1 (West & Harrison, 1997). The discount factor was optimized separately for the univariate and multivariate DLM by trying values of δ between 0 and 1 by steps of 0.01.

Forecasts of the expected EC values at time step *t* are made according to the observation equation:

$$\boldsymbol{Y}_t = \boldsymbol{\theta}_t + v_t \quad (5)$$

, where $\boldsymbol{Y}_t$ is the observation vector with a length of 1 for univariate model and 4 for the multivariate model. The error term is defined as $v_t \approx N(\underline{0}, \boldsymbol{V})$ with $\boldsymbol{V}$ being the observational co-variance matrix, with the dimensions 1x1 for the univariate model and 4x4 for the multivariate model. The values of $\boldsymbol{V}$ were found using the expectation maximization algorithm, as described in detail by West & Harrison (1997).

At each observation time, the forecast errors were calculated according to eq. 6.

$$\boldsymbol{e}_t = \boldsymbol{Y}_t - \boldsymbol{\theta}_t \quad (6)$$

Given the forecast errors for each observation time, the parameter vector values are updated using the Kalman filter (West and Harrison, 1997). The forecast variance-covariance matrix is estimated as part of the Kalman filtering (West and Harrison, 1997). The dimensions of this matrix is 1x1 for the univariate DLM and 4x4 for the multivariate DLM.

Standardization

For the MA, EWMA, and the univariate DLM, the forecast errors were standardized using the forecast variance, according to eq. 7.

$$u_t = \frac{e_t}{\sqrt{\sigma_{z_t}^2}} \quad (7)$$

For the multivariate DLM, the forecast errors were standardized in the same way, except using only the diagonal values of the forecast variance-covariance matrix.

Observation classification

The four optimized time series filtering models were applied to the of EC observations in the original training set, resulting in four different new training sets. These new training sets contained 12 predictor variables per observation, namely the unfiltered EC values, the filtered EC values, and the standardized forecast errors for each of the four quarters. Based on these predictor variables, different machine learning methods were trained to classify the individual milkings as being from a mastitis positive or negative milking, as described below.

Random Forest and Bayesian Network

We set-up an experiment in the machine learning and data mining tool WEKA (Witten and Frank, 2005) to be used for each of the four training datasets. In the experiment, two main algorithms were selected, namely Random Forest (RF) and Bayesian network (BN). Each of these two algorithms were set up with different combinations of parameter settings resulting in a total of 31 different model configurations to be evaluated (for details, see Tables 1 and 2). Each of these model configuration was evaluated with 5-fold cross validation on each of the training datasets. Mastitis (yes/no) was used as the categorical output variable. No further pre-processing was performed.

Table 1: Overview of the parameter settings used with the random forest method

Main algorithm	Parameter settings		Abbreviation
	No. of trees	Seed	
Random Forest	3	1 - 5	RF031 - RF035
	10	1 - 5	RF101 - RF105
	25	1 - 5	RF251 - RF255

Table 2: Table 3: Overview of the parameter settings used with the Bayesian network method

Main algorithm	Parameter settings		Abbreviation
	Search Algorithm	Max no. of parents	
Bayesian Network	Local	1 - 5	BNL1 - BNL5
	TAN	N/A	BNTan
	Tabu	1 - 5	BNT1 - BNT5
	Hill Climber	1 - 5	BNH1 - BNH5

Each of the 31 model configurations produced a probability of having mastitis for each record in each of the four training sets. If this probability was > 0.5, the final output was categorized as 1 (mastitis predicted), and else the final output was categorized as 0 (no mastitis predicted). These predictions were compared with the true status of each record to assess true positive (TP), false negative (FN), true negative (TN), and false positive (FP) predictions per records. Per model, threshold settings were changed such that Se of finding ME was ~60% at milking level. At that level, FAR1000 was reported. This process was repeated for each of the four training sets using Weka experimenter. Based on these performance parameters, one model configuration for each of the two main algorithms was selected for further analyses.

Model testing

The selected configurations of the three classification models (RF, BN, and Sewhart control chart) were applied to each the four related test dataset, probabilities were produced for each records, and these probability were transformed into a 0/1 output, depending on the threshold. We then applied the time-window proposed by Kamphuis et al. (2016) to compute TP and FN alerts for each ME. This time-window assumes that an alert from any mastitis detection model can be expected from up to two milkings prior to the first milking of a ME, and then for the entire duration of a ME. Alerts earlier than the two milkings prior to the start of an ME, or after the last milking of an ME are thus considered as a FP alert. Subsequently, each ME is counted as either being missed (one FN alert) or as one correctly identified ME (one TP alert), although more than one milking within a ME could have received a TP alert. For milkings not belonging to a ME, which thus had a true no-mastitis status, no time-window was used. This means that each no-mastitis milking receiving an alert by the model were counted as FP, and each no-mastitis milking not receiving an alert were counted as TN. The TP and FN were used to compute Se, while the TN and FP alerts were used to compute the number of alerts per 1,000 milkings (FAR1000). Per model, threshold settings were adjusted such that Se of finding ME was ~60 %. At that level, FAR1000 was reported. Furthermore, the specificities and error rates were calculated as secondary measures of performance for the different classification methods.

Results and Discussion

The aim of this study was to compare the performances of different combinations of filtering and classification methods under standardized conditions. The ability of human milkers to detect clinical mastitis has been reported with an average sensitivity of 80 %, although this number in practice depends on the skill of the milker and the severity of the mastitis case (Hillerton and Kliem, 2002). For comparison, the scientific literature reports an average sensitivity of 60 % when using automated detection systems in the field (Hogeveen *et al.*, 2010). For this reason the sensitivity in this study was fixed at 60%, and thus the threshold for which outputs would count as alarms was optimised for each classification methods. Table 3 summarizes the performances achieved with the various method combinations.

Table 3: Preliminary results of the data driven methods

Classification method	Filtering method[1]	Sensitivity (%)	Specificity (%)	FAR1000[2]	Error rate (%)
Random Forest	MA	59.9	74.8	217.7	95.6
	EWMA	60.8	73.7	227.6	95.8
	DLM uni	60.2	75.2	214.2	95.6
	DLM multi	60.7	74.1	224.2	95.7
Bayesian Network	MA	60.5	73.0	233.3	96.0
	EWMA	60.3	73.8	226.5	95.9
	DLM uni	59.4	71.8	234.7	96.1
	DLM multi	59.9	74.1	224.5	95.8

[1] Filtering methods: moving average (MA), exponentially weighted moving average (EWMA), univariate dynamic linear model (DLM uni), and multivariate DLM (DLM multi).
[2] Number of alerts per 1,000 milkings

The results obtained for the test data were comparable to the results for the training data, which argues for the validity of the model and suggests that the model does not over-fit to the data. This may indicate that the model is generally applicable. Table 3 shows the detection performance of the classification methods using different filtered datasets. The Sp is ranging between 71% and 75%, with a Se of 60%. The performances of the models in this study are lower compared to findings form the literature. This, however, was expected since we do not search for the best possible model, but rather seek to compare different filtering and classification methods more objectively. The specificity rate obtained with the classification seem rather similar. The error rates were high, ranging between 95% and 96%. Since there are many more days with a healthy stage than days of mastitis, it causes a greater likelihood for FP to arise, which has an impact on the error rate. The results of the filtering methods showed similar results.

The different filtering methods do not vary a lot in Sp, FAR1000 and error rate. The number of CM episodes indicated with Random Forest suggest to be higher than for the Bayesian Network. The FAR1000, however, indicates that more alert are generated with Bayesian Network than for the Random Forest. Despite some differences the results suggest that the performances of these two classification methods do not vary a lot. In the next step of this study we will look also at the relative more straightforward methods like a Shewart Control chart. We hypothesis that the performances of such methods are less compared to the more advanced classification methods. Beside, we know that the Shewart Control works fine with continuous variables like EC, but utilizing the categorical variables (e.g. parity), this simple method will quickly become much more complicated to implement.

The results of this study should be interpreted as a relative comparison. According to Hamann and Zecconi (1998) using EC in milk as a mastitis is a good indicator. However, the performances of the classification methods are expected to be improved when including other variables, like milk yield and SCC. Additionally, including historic data (from e.g., previous milkings) are also expected to improve the detection performances.

The naïve approach we used in this study was necessary to enable a fair comparison between the performances of classification and filtering methods.

Conclusions

This study aimed at evaluating performances of several data driven methods for the early detection of mastitis, using similar conditions (i.e. data input, data selection criteria, time-window and gold standard). So far, there is an indication that our naive approach of data handling results in no clear distinction in performance between the different methods.

References

De Mol, R.M., and Ouweltjes, W (2001) Detection model for mastitis in cows milked in an automatic milking system. *Preventive Veterinary medicine* 49, 71-82.

Dominiak, K.N., and Kristensen, A.R. (2017) Prioritizing alarms from sensor-based detection models in livestock production – a review on model performance and alarm reducing methods. *Computer Electronics in Agriculture* 133, 46-67.

Ebrahimie, E., Ebrahimi, F., Ebrahimi, M., Tomlinson, S., and Petrovski, K.R., 2018. Hierarchical pattern recognition in milking parameters predicts mastitis prevalence. *Computers and Electronics in Agriculture* 147, 6-11.

Hogeveen, H., Kamphuis, C., Steeneveld, W., and Mollenhorst, H. (2010) Sensors and Clinical Mastitis—The Quest for the Perfect Alert. *Sensors* 10, 7991-8009.

Hillerton, J.E., Kliem, K.E. (2002) Effective treatment of Streptococcus uberis clinical mastitis to minimize the use of antibiotics. *Journal of Dairy Sciences* 85, 1009–1014.

Kamphuis, C., Dela Rue, B.T., and Eastwood, C.R. (2016). Field validation of protocols developed to evaluate in-line mastitis detection systems. *Journal of Dairy Sciences* 99, 1619-1631.

Khatun, M., Clark, C.E.F, Lyons, N.A., Thomson, P.C., Kerrisk, K.L., and García, S.C., (2017) Early detection of clinical mastitis from electrical conductivity data in an automatic milking system. *Animal Production Science* 57, 1226-1232.

Schepers, A.J., T.J.G.M. Lam, Y.H. Schukken, J.B.M. Wilmink, and W.J.A. Hanekamp. 1997. Estimation of Variance Components for Somatic Cell Counts to Determine Thresholds for Uninfected Quarters. J. Dairy Sci. 80:1833–1840.

West, M., and Harrison, J., 1997. Bayesian Forecasting and Dynamic Models, 2nd ed. Springer, New York, USA.

Witten, I.H., and Frank, E. (2005). *Data Mining; Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Publishers, San Francisco.