Analysis of deviations of reticuloruminal pH data of dairy cattle and development of a rumen condition scoring system

Contents

Current situation (problem description) and aim of the project

Due to the technical development, reticuloruminal pH data in cattle can be continuously accessed and monitored through remote sensors. (boluses) These data are then used to infer that given cow suffers in subacute ruminal acidosis or in any other health problems, which could lead in the future to a considerable decrease in the milk yield. The industry partner involved in the Capstone Project develops and produces such boluses and sells them to farm owners and farm managers. Provision of an online tool (Dashboard) is also part of the service, where pH data can be monitored, however to detect abnormalities in pH pattern(s) simple threshold value(s) are used: shall a pH value exceed or fall behind given threshold(s), this must be considered as an alert. Even though this "threshold" approach is getting more and more sophisticated, (usage of daily mean(s) by animals or by farms, etc.) it is a research topic with increasing importance in this segment of agriculture to apply more sophisticated time series models in order to describe the behaviour and fluctuation of pH observations.

My Capstone Project is aimed at describing this fluctuation with time series models, but also goes farther in two aspects: firstly I used always three models to capture patterns in the data enabling also the comparison between them through fit measures, (developed methodology offers the possibility of the inclusion of three models with parametrizable features) and secondly – due to the complexity and difficult interpretability of the results of the models – I established a simplified scoring system, which will show farm owners deviations from the established pattern(s) by means of scores characteristic for given animal.

The applied approach throughout the whole project is data analytics, and not veterinarian or chemical: this means that developed flexible methodology offers the possibility of choosing between models and to adjust parameters of the scoring system, but does not take the responsibility of studying the impact of lower/higher pH values or deviations on the health of animals.

Data

To carry out the project, data of 5 farms were used. The number of used boluses by farm varies between 6 and 15. Any bolus transmits pH data in every 10 mins, lifespan of a given bolus is approx. 90 days, but in some cases data were sent on the 170th day also. This "dimension" of the data means, that practically all important clients' data of the industry partner were involved in the project.

Provided (historical) data were used to fit time series models, but also to imitate data flow and business processes at the industry partner to establish the required scoring system which would also work in the real business.

Activities carried out in the framework of the project

The completion of the project and the achievement of stated goals required following activities to be performed:

- Data cleaning
- Fitting time series models
- Generating forecast(s) based on time series models
- Compare forecasts and real observations and take consequences of the differences (Scoring system)

Data cleaning

As a first step, it was necessary to order and clean raw data plus to fill in missing values. Provided data showed different symptoms of untidiness, for example entire days missing at the end or after the lifespan of the bolus, lack of all signals on hourly level, observations transmitted twice and extreme values. All these problems shall be considered as normal in data analytics, however the peculiarity of missing observation value and corresponding time stamp together enforced the creation of a respective time series with adequate periodicity of time stamps, merge with available data and fill in missing values.

Major gaps at the end (after the lifespan of the bolus) were simply cut. All these transformations were necessary, because both the model fitting and forecast(s) generating processes are very sensible for any type of untidiness in the data.

Fitting time series models

The goal of this section is to capture and describe the fluctuation of pH observations with (time series) models. As my project builds on a former study in the same topic, results of the previous research – namely a sine waves model fitted on pH observations – were used in the form of code and parameters establishing one of three models in my project.

To extend pattern discovery possibilities, project's methodology offers the inclusion of two alternative AR(I)MA models also aimed at capturing fluctuation in the data. Parameters of AR(I)MA models are also adjustable, (bolus by bolus) and final choice which model to use is up to the partner/user.

Generating forecast(s) based on time series models

After fitting the models, the biggest challenge was how to use historical data for forecasting for the future and how to imitate business processes in order to elaborate useable codes for real production at the same time.

Considering the real life scenario, at given point in the time (where time means the whole lifespan of the sensor/bolus) not all observations transmitted during the whole lifecycle of the bolus are known, just part of it: exactly what happened before that point in the time. As a consequence, only this group of observations can be used to determine the "ideal" (forecasted) value, which is then compared to the observed value.

The used technique for this imitation is "time series cross-validation", which served above stated complex goals: the partitioning of historical data into two sections makes it possible to “train” the model on one part of observations and forecast for to other part and then compare forecasted values to real observations in this second group. (in my approach the size of the second group is always 1 observation, and this “traditional” application of time series cross-validation allowed the calculation of forecast accuracy measures to compare models also) Simultaneously this process maps data flow and processing in the business: there are given number of observations available from the past, based on this a forecast is generated for the next 10^{th} mins observation, which is then compared with the real observation coming into the system. In the next step given number + 1 observations are used for forecasting the next 10^{th} mins observation and so on.

These computations (called one-period-ahead-forecast) were performed for all three models in case of each bolus to enable the choice between models. (which choice is not necessarily based on performance of fit measures) The method ordered to each real pH observation (after the last element of the “training” part) three forecasts and three errors (difference between forecast and real observation) generated according to the three time series models applied. These differences were evaluated by the scoring system.

Compare forecasts and real observations and take consequences of the differences – the Scoring System

The scoring system is flexible in the sense that any forecasts and corresponding errors of the three models (and not only the best performing) can be used for computations. The main goal of the scoring system (function) is to construct tolerance bands around the forecasted values and evaluate in which tolerance band real observation falls in order to take consequences in form of score points. If five percentage values (for example 1%, 2%, 3%, 4%, 5% in tolerance argument of the function) are given as tolerance bands indicators, and the observation falls within the closest band (1%) to the forecast, then score point will be 5, while if it's also outside of the farthest tolerance band (5%) then the score point will be 0. Technically the function evaluates the coincidence of each real observation within each tolerance bands constructed around the corresponding forecast, (obviously if observation falls within the closest tolerance band, then it will fall within the farthest also) and calculates the number of cases when coincidence criterion is satisfied. The system accepts up to ten tolerance bands, depending on partner/user’s choice to which extent distinguish results of the model.

The whole system performed well when testing on different bolus data. Obviously, the chosen time series model plus the number and scale of applied tolerance bands have a great impact on the results: a time series model with good fit measures and broad tolerance bands doesn't lead to "informative" score values, while the sine waves model (less fitting) with narrow tolerance bands delivered distinguishable score points for pH observations.

Above example shows that choices about input parameters at every step of the process matter, therefore my goal was to ensure the flexibility and thus the emergence of not only statistical/analytical but also veterinarian or chemical aspects. **The final product offers not only the criterion of flexibility but also the possibility of direct application at the industry partner.**