Computers and Electronics in Agriculture 147 (2018) 6–11

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

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

Original papers

Hierarchical pattern recognition in milking parameters predicts mastitis prevalence

Esmaeil Ebrahimie[a,b,c,d,*], Faezeh Ebrahimi[e], Mansour Ebrahimi[e], Sarah Tomlinson[f], Kiro R. Petrovski[f,*]

[a] *Adelaide Medical School, The University of Adelaide, Adelaide 5000, Australia*
[b] *Division of Information Technology, Engineering & Environment, School of Information Technology and Mathematical Sciences, University of South Australia, Adelaide, Australia*
[c] *Faculty of Science and Engineering, School of Biological Sciences, Flinders University, Adelaide, Australia*
[d] *Institute of Biotechnology, Shiraz University, Shiraz, Iran*
[e] *Department of Biology, University of Qom, Qom, Iran*
[f] *School of Animal and Veterinary Sciences, The University of Adelaide, Adelaide 5371, Australia*

ARTICLE INFO

Keywords:
Machine learning
Mastitis
Milking parameters
Random Forest model
Somatic Cell Count

ABSTRACT

The aim of this study was to develop a predictive model for mastitis incidence, independent from Somatic Cell Count (SCC), to provide an alternative, simple, and cost-effective approach for mastitis risk management based on available milking parameters. The test-day Somatic Cell Count (SCC) is the most common indicator for Sub-Clinical Mastitis (SCM) surveillance in dairy industries worldwide. However, SCC is highly variable between days, raising major concerns for its reliability. This caveat highlights the need for longitudinal/frequent monitoring of SCC and/or developing alternative approaches for SCM surveillance. A considerable proportion of available milking data such as Milk Volume, Protein, Lactose, Electrical Conductivity (EC), Milking Time, and Peak Flow provide the possibility of pattern recognition and model discovery towards mastitis occurrence. Developing a predictive model involves: (1) finding the threshold (cutoff) of different predictive milking parameters and (2) finding the best combination of features that lead to mastitis and their hierarchical pattern/order. Here, in a large-scale study on 346,248 milking records, for the first time, we evaluated four different decision tree algorithms (Decision Tree, Stump Decision Tree, Parallel Decision Tree and Random Forest Decision Tree) with four different criteria (Accuracy, Info Gain, Gini Index and Gain Ratio) run on 11 datasets (original dataset and 10 created datasets by attribute weighting selection algorithms). Therefore, 572 models were evaluated and compared by 10-fold cross validation. The performance of each decision tree in drawing an inverted tree; with the most important feature at the root and less important variables as the leaf; was calculated by 10-fold cross validation. Random Forest Decision Tree with Gini Index criterion was the best model for predicting mastitis from milking parameters with a high accuracy of 90%. Decision Tree models identified a strong pattern for SCM in milking data where all (100%) of cows with low levels of lactose (Lactose ≤ 4.5 g/L) and low milk volume (Volume ≤ 21.7 L) had mastitis. In addition, a significant pattern was found for identifying healthy cows by high levels of lactose (Lactose ≥ 4.5 g/L) and low levels of EC (EC ≤ 5.2). This study doccuments that milking parameters mined by the Decision Tree Random Forest model can be utilised to accurately predict SCM. The findings can be employed to increase the reliability of test-day SCC or as SCC-independent and cost-effective predictors of SCM.

1. Introduction

Mastitis is a costly disease with detrimental economic impacts on the dairy industry (DeGraves and Fetrow, 1993; Hogeveen et al., 2011). Worldwide, it is estimated that the economic losses of mastitis range from €61 to €97 per cow on a farm (Hogeveen et al., 2011). It is necessary to develop accurate surveillance methods for determining the incidence and prevalence of mastitis in dairy herds. Mastitis occurs in two forms: clinical (CM) and Sub-Clinical Mastitis (SCM). SCM is approximately 40 times more common than CM, and is difficult to detect because clinical signs are not apparent in the infected cow.

SCM prevalence is commonly determined by the number of somatic

[*] Corresponding authors at: School of Animal and Veterinary Sciences, The University of Adelaide, Roseworthy Campus, Mudla Wirra Rd, Roseworthy 5371, Australia.
E-mail addresses: esmaeil.ebrahimie@adelaide.edu.au (E. Ebrahimie), kiro.petrovski@adelaide.edu.au (K.R. Petrovski).

https://doi.org/10.1016/j.compag.2018.02.003
Received 25 June 2017; Received in revised form 4 January 2018; Accepted 4 February 2018

cells in stripped milk, called Somatic Cell Count (SCC). SCC tests reflect the inflammation status of bovine mammary glands at the time of sampling; mainly caused by bacterial infection. For the estimated prevalence to be meaningful, SCC must be sufficiently stable between test days. SCC stability depends on many variables such as the farm's pathogen prevalence profile. The sensitivity and specificity of a SCC test in determining the threshold of real infection may be hampered by a high prevalence of minor pathogens. Some studies have already highlighted the importance of these factors (Goncalves et al., 2016; Petrovski et al., 2006). Altogether, SCC is highly variable between days, raising major concerns for its reliability. This issue becomes more important in light of the economic losses due to SCM, highlighting the urgent need for new assessment tools.

An alternative approach for predicting SCM incidence can now be investigated in milking parameters since automated monitoring devices have enabled milk samples to be analysed for a range of production parameters, such as the percent of fat, lactose or protein and milk volume. The quality and quantity of milk are also affected in SCM, showing signs such as small flakes, clots, and a watery appearance (Goncalves et al., 2016; Petrovski et al., 2006). The availability of a considerable amount of data on milking quality and quantity, encouraged us to investigate whether these parameters are stable enough to be utilised to develop predictive models of SCM, independent of SCC, using data mining tools.

Recently, data mining based on machine learning has emerged as the method of choice when large sets of data need to be processed and refined (Ebrahimi et al., 2014; Kargarfard et al., 2015, 2016). Supervised machine learning is the process of extracting implicit, previously unknown, and potentially useful information from data. It can also be defined as data analysis for finding regularities and patterns in a given dataset. While machine learning can be defined as an extension of multivariate on large datasets, it has a better performance compared with multivariate methods previously employed in various fields, including medicine, biology, finance etc. (Ebrahimi et al., 2010, 2011; Bharat et al., 2016). Machine learning is extensively used in biological studies because of its pattern recognition abilities (Hande Küçükönder et al., 2015; KayvanJoo et al., 2014). In order to identify the right combination of features and thresholds that are able to predict SCM from milking parameters, it is necessary to find the machine learning algorithms with high performance. A few studies have employed data mining in studying mastitis sensor data from automatic milking systems, mainly focusing on multivariate analysis (Kamphuis et al., 2008; Ortiz-Pelaez and Pfeiffer, 2008).

Decision trees are one of the best methods known for classification in machine learning algorithms (Ebrahimi et al., 2014). Due to their simple structure, they are easily assimilated by humans, easily constructed, and superior to other known methods of classification (Gehrke et al., 1998). Decision trees are a visual representation of data where an attribute is classified by its relation to other attributes. We have applied this algorithm on our dataset in order to determine which features play a significant role in detecting SCM, and then to establish a SCC-independent model for SCM.

2. Materials and methods

2.1. Data collection

Data was gathered in a 2 years period (July 2011 to June 2013) from a commercial New Zealand dairy farm in Ongaonga, Hawkes Bay. Milking was performed twice a day on 2400 mixed-age Holstein Friesian cows. An electronic automated monitoring system was used to measure features of harvested milk, including volume, weight, fat, protein, lactose, electrical conductivity (EC), milking time, peak flow and SCC. The SCC was calculated by an on-line detector, CellSense®, which measures SCC one minute after milking as described previously by (Meijering et al., 2004). Data cleansing and preparation was performed to eliminate incomplete records or those with recording errors before analysis.

2.2. Preparation of dataset

The original dataset had 346,248 recorded samples (or rows) with eight variables [Volume (Vol), Fat, Protein (Prot), Lactose (Lact), EC, Milking Time, Peak Flow and SCC)] which were recorded at each milking time. Cow records were classified into two groups, based on Australian based on the Australian definition of mastitis (≥250,000 cells/mL) as non-mastitis group and mastitis group.

The dataset was imported into Rapid Miner software (RapidMiner 5.0.001, Rapid-I GmbH, Stochumer Str. 475, 44,227 Dortmund, Germany) and SCC variable was set as output (label or target) variable, and other variables as dependent inputs.

2.3. Attribute weighting and attribute-selection

As described previously, the most important variables contributing to sub-clinical mastitis were determined by ten attribute weighting models (Ebrahimi et al., 2011). The algorithms were: Information Gain, Information Gain Ratio, Rule, Deviation, Chi-squared statistic, Gini Index, Uncertainty, Relief, Support vector machine and Principal component analysis. As data was normalised before running the models, all weights were calculated as values between 0 and 1; showing the importance of each attribute to the target attribute. To provide the possibility of comparing results of different normalised attribute weighting models, as previously described (Ebrahimi et al., 2014), an index based on the intersection (agreement) of different weighting models was created, and important features were selected to be key indicators of mastitis. These important features have been defined by most weighting models (weight cut-off of 0.5, 0.75, or > 0.95).

Subsequent to running the aforementioned attribute weighting models on the original balanced dataset (BDS), 10 new datasets were created based on all the variables with weight value more than 0.5. Newly created datasets were given their attribute weighting algorithm names (Information Gain, Information Gain Ratio, Rule, Deviation, Chi Squared, Gini Index, Uncertainty, Relief, SVM and PCA).

2.4. Decision tree algorithms

A decision tree is an inverted tree-like graph which has a root at the top and grows downwards; representing data for easy interpretation. The main goal of decision tree models is to create a classification model that predicts the value of label/target class (here SCC variable with two class of Yes and No) based on several input features or variables (here Lact, Vol, EC, Milking-time, Fat, Prot and Peak-flow). Each interior node of the tree corresponds to one of the input attributes. The number of edges of a nominal interior node is equal to the number of possible values of the corresponding input attribute.

Various tree induction algorithms were used in this study, briefly they were:

Decision Tree model: This model uses recursive partitioning which repeatedly splits the values of attributes.

Decision Stump model: This operator learns a decision tree with only one single split. This operator can be applied on both nominal and numerical datasets.

Random Forest model: This operator generates a set of a specified number of random trees i.e. it generates a random forest. The resulting model is a voting model of all the trees.

Random Tree model: This operator uses only a random subset of attributes for each split.

Sixteen decision trees were employed on the original dataset (BDS) and 10 new datasets were created by attribute weighting selection algorithms. Decision tree models were: Decision Tree, Decision Tree Parallel, Decision Stump and Random Forest followed by four different