Assessment of the future productivity level of dairy cows: between dream and reality

M. Bovo[1], L. Ozella[2], E. Fiorilla[2] and C. Forte[2]
[1]Department of Agricultural and Food Sciences, University of Bologna, Viale Fanin 48, Bologna (BO), Italy
[2]Department of Veterinary Sciences, University of Turin, Largo Paolo Braccini 2, Grugliasco (TO), Italy

marco.bovo@unibo.it

Abstract

The challenge of increasing the economic and environmental sustainability of the dairy cattle sector involves several factors like milk yield and milk quality levels, cow health and wellbeing, efficient resource use, and emissions reduction. Among the different features, the milk production for lactation, throughout the productive career of a cow, is perhaps the parameter that all farmers would like to know for a more efficient planning of entries and exits from the herd. In fact, if one the one hand numerous researches have studied and are still addressing the problem from a genetic point of view, on the other hand, few studies have focused on the definition of tools for predicting the productivity class future lactations of cow. In the study, firstly two supervised learning methods, i.e., Super Vector Machine and K-Nearest Neighbors, have been applied to a large dataset of 720 complete lactations, with the object to train machine learning tools for the classification between first and second lactation. Then, for those cows having available the data of first and second lactation curve, the two classification methods have been trained and tested for the attribution of the second lactation productivity level (i.e., low, medium or high) starting from the data of the first lactation. The classification methods reached accuracy values ranging from 70% to 73%. These values seem very encouraging and indicate that the predictors selected, despite their simplicity, look very promising and could pave the way for the definition of enhanced future models.

Keywords: KNN, SVM, PLF, dairy cow, AMS, big data

Introduction

The challenges of the sustainability in the dairy cattle sector involve milk yield and milk quality levels, cow health and wellbeing, efficient resource use, and emissions reduction (Strpić et al., 2020). Due to the effects on milk production and quality, which have an impact on how effectively natural resources are used, animal welfare is, at the end, directly linked to sustainability, and as widely demonstrated, increasing animal welfare usually increase the milk yield (Allen et al., 2015; Kino et al., 2019). To this regard, in the recent years, several steps forward have been made to increase the production per lactation of the individual animal by working on the genetic selection, on feeding, increasing animal welfare and the quality of the housing environment (Chamberlain et al., 2022; Zhou et al., 2022). Many of these actions are related to daily management decisions that the farmer, nowadays, can undertake with the help of commercial management tools or decision-support systems often associated with sensors or technologies that allow real-time monitoring of the production and health status of the individual animal (Giannone et al., 2023). In fact, following the Precision Livestock Farming (PLF) approach (Berckmans, 2014; Tullo, Finzi and Guarino, 2019; Lovarelli, Bacenetti and Guarino, 2020), in technological farms, data concerning different parameters of behavior and activity of cows, animal health and welfare are collected from different sensors (e.g., individual cow data recording system, activity tags such as pedometers or neck collars, ear tags for rumination monitoring, automatic concentrate feeders), and used for the daily management of the herd (Bovo et al., 2020). Furthermore, the growing widespread of automatic milking systems (AMSs) and electronics milking parlors (EMPs) provide farmers and technicians with continuous series of detailed data useful to assess health conditions and evaluate parameters connected to the milk quality and quantity (Ozella et al., 2023). But, while most of the recent studies investigated models primarily focusing on the prediction of the daily milk yield for the running lactation period (Jones, 1997; Ji et al., 2022) one of the still open matters involves the prediction of cow productivity in future lactation periods (Rebuli et al., 2023). This aspect is particularly important in the first years of life of cows, because as is well known, the first lactation usually has a lower production compared to subsequent lactations (Masía et al., 2020), and for a farmer it is important to know, as soon as possible if, compared to the other animals in the herd, a specific animal will have, on a long term, high, medium or low milk productivity (Arulnathan et al., 2020). Then, for the dairy sector, one of the next big challenges will be the development of a system/tool able to classify the future productivity of a cow based on what the cow produced in the past or is currently producing (Bovo et al., 2024; Giannone et al., 2023). With reference to this, nowadays, the availability of large dataset collected by AMSs and EMPs, make possible the application of big data approaches (Fuentes et al., 2020) and especially those based on machine learning algorithms (Dulhare, Ahmad and Ahmad, 2020). For these research problems, a classification learner can represent one of the most promising numerical tools (Frades and Matthiesen, 2010; Everitt et al., 2011). Actually, the problem could be divided into two closely related aspects. The first is related to the identification of the features that characterize the lactation number of an animal, classifying first lactation separated from the following lactations. All this allows to estimate the value of a metric providing a measure of the distance between the two clusters (i.e., the cluster of the daily milk yield time series of the first lactation and the cluster of the time series of the lactation periods two and higher). Instead, the second research aim is related to the classification of the productivity level (e.g., low, medium or high) of a cow, in future lactations, starting from the knowledge of the productivity features of its first lactation.

In the present paper these two still open questions have been approached starting from the assumption that first and second lactations have milk yield trends considerably different. So, two supervised learning methods, i.e., the Super Vector Machine (SVM) and the K-Nearest Neighbors (KNN), have been applied to a large dataset with the object to train models for the classification between first and second lactation curves. Finally, for those cows having available data from first and second lactation curve, the two learning methods have been used for the attribution of a productivity class (i.e., low, medium, high) for the second lactation of a cow starting from data of the first lactation of the same cow.

Material and methods

Dataset description

The dataset used in the work was gathered from March 2020 to May 2022 in 13 farms located in the Po Valley region, in northern Italy. The 13 farms are equipped with a Merlin AMS (Fullwood Packo, England) that collects data on daily milk yield (DMY) and milk quality (i.e. fat, protein and lactose) for each cow. The size of the 13 herds is similar and the farms have about 60 milking cows each one. In order to uniform the dataset length of the different cows, the lactation period was assumed at the maximum of 305 days in milk (DIM). Therefore, for cows having a long lactation, only the first 305 days have been considered. Moreover, the lactation data has been considered valid for the analyses only if it contains data for at least 250 days (Perez Garcia et al., 2023, 2024). In total, the number of unique animals in the dataset is 683 and the number of valid lactations is 720 (i.e., 465 first lactations and 255 second lactations).

Classification methods

In this section the two machine learning algorithms used for the lactation classification are described. For both classifiers the k-fold cross-validation procedure was followed in order to keep a stable training-test configuration and to have the best estimation of the classification performance (Witten, Frank and Hall, 2011). The confusion matrix is the tool we have used to validate the accuracy of the classification methods.

SVM algorithm

The support-vector machine (SVM) algorithm developed by Vapnik (Cortes and Vapnik, 1995) is based on statistical learning theory. At the first approximation SVM finds a separating line (or hyperplane) between data of different classes. SVM is an algorithm that takes the data as an input and outputs a line that separates those classes. According to the SVM algorithm it finds the points closest to the line from both the classes. These points are called support vectors. Now, we compute the distance between the line and the support vectors. This distance is called margin. The goal of the algorithm is to maximize the margin in order to define the optimal hyperplane (i.e. the hyperplane for which the margin is maximum). If data are clearly not linearly separable it's impossible to draw a straight line that classify the data and then SVM can convert original data to linearly separable data with a nonlinear transformation of the original space. In its most simple type, SVM doesn't support multiclass classification natively. It supports binary classification and separating data points into two classes. For multiclass classification, the same principle is utilized after breaking down the multiclassification problem into multiple binary classification problems.

KNN algorithm

The k-nearest neighbors (KNN) algorithm (Cover and Hart, 1967) is a supervised machine learning algorithm used to solve both classification and regression problems. The algorithm assumes that similar things exist in close proximity. In other words, similar things are sufficiently near to each other. The main steps followed by KNN are the following:

1. Initialize K to your chosen number of neighbors;
2. For each example in the data, it calculates the distance between the query example and the current example from the data and then adds the distance and the index of the example to an ordered collection;
3. Sort the ordered collection of distances and indices from smallest to largest by the distances;
4. Pick the first K entries from the sorted collection;
5. Get the labels of the selected K entries;
6. If regression, return the mean of the K labels, otherwise if classification, return the mode of the K labels.

Results and Discussion

Performance of the classification first/second lactation

For the purpose of the first research problem, the two classificatory algorithms i.e., SVM and KNN were applied to the 305-day time series of the 720 lactation curves in the dataset. The classification accuracies are equal to 0.79±0.03 and 0.88±0.05 respectively for SVM and KNN. The two confusion matrices are, respectively, equal to [402, 63; 72, 183] and [444, 21; 63, 192]. From the analysis of the results, it emerges that KNN has higher accuracy in the classification. In fact, 585 lactations out of 720 were correctly labelled in comparison to the 507 of the SVM method. The results of the classification procedure confirm the existence, in the time series structure, of some intrinsic characteristics used by the classification methods adopted here.

Assessment of the productivity class of future lactations

As already discussed in the previous sections, in the dairy sector, one of the main challenges is the definition of a model evaluating the future productivity of a cow (in terms of milk yield) based on milk yield of previous lactations. So, in this work, for those cows having available data from first and second lactation curve, two learning methods have been used for the attribution of the productivity class (i.e., high production (HP), medium production (MP), low production (LP)) of the second lactation of a cow starting from data of the first lactation. The dataset contains the data of first and second lactation of 37 different cows. In the first lactation, 13 has HP class, 9 has MP class and 15 LP class. The tools selected to approach this task are the classification methods used before, i.e., SVM and KNN algorithms. It is worth to note that only a partial group of cows maintains the same productivity class moving from first to second lactation. In fact, 24 out of 37 lactations (i.e., 65%) maintained the same label whereas 35% of the lactations have moved to another productivity class. Further future research will investigate on a larger dataset the stability of this percentage, but the value is in accordance with the outcomes in (Rebuli et al., 2023) confirming as an important group of primiparous cows has not stable production level moving from first to the subsequent lactations.

The classification methods reached an accuracy of 73% (i.e. (10+17)/37×100) and 70% (i.e. (10+16)/37×100) respectively for SVM and KNN. In the opinion of the authors these first values are very encouraging and indicate that the selected features, despite their simplicity, look very promising and entail further studies and development. It seems worth noting that the method could become particularly interesting for practical applications, as it represents a viable support-to-decision tool that farmers can adopt for the selection of the most productive animals to be kept in the herd compared to those to be discarded.

Conclusions

In this paper two supervised learning methods have been trained and tested with the main aim to properly recognize and label the first and the second lactation curves of dairy cows. The two classification algorithms have been applied to the raw dataset and then after the application of four different dimensionality reduction methods. Finally, for those cows having available data from first and second lactation curve, the two classification methods have been used for the attribution of the productivity class (i.e., low, medium, high) of the second lactation of a cow starting from its data on the first lactation. The classification methods reached accuracy values ranging from 70% to 73%. These values seem very encouraging and entail further studies for the definition of enhanced models useful as decision support tools for farmers for early selection of the most productive animals to kept in the herd or discharge.

References

Ahmad, N. and Nassif, A.B. (2022) 'Dimensionality Reduction: Challenges and Solutions', ITM Web of Conferences [Preprint]. Available at: https://api.semanticscholar.org/CorpusID:247463581.

Allen, J.D. et al. (2015) 'Effect of core body temperature, time of day, and climate conditions on behavioral patterns of lactating dairy cows experiencing mild to moderate heat stress', Journal of Dairy Science, 98(1), pp. 118–127. Available at: https://doi.org/10.3168/jds.2013-7704.

Arulnathan, V. et al. (2020) 'Farm-level decision support tools: A review of methodological choices and their consistency with principles of sustainability assessment', Journal of Cleaner Production, 256, p. 120410. Available at: https://doi.org/https://doi.org/10.1016/j.jclepro.2020.120410.

Berckmans, D. (2014) 'Precision livestock farming technologies for welfare management in intensive livestock systems', Rev. sci. tech. Off. int. Epiz, 33(1), pp. 189–196. Available at: https://doi.org/10.20506/rst.33.1.2273.

Bovo, M., Benni, S., Barbaresi, A., Santolini, E., Agrusti, M., Torreggiani, D., & Tassinari, P. (2020). A Smart Monitoring System for a Future Smarter Dairy Farming. 2020 IEEE International Workshop on Metrology for Agriculture and Forestry, MetroAgriFor2020 - Proceedings, 165–169. https://doi.org/10.1109/MetroAgriFor50201.2020.9277547

Bovo, M., Ceccarelli, M., Agrusti, M., Torreggiani, D., & Tassinari, P. (2024). DAIRY CHAOS: Data driven Approach Identifying daiRY Cows affected by HeAt lOad Stress. Computers and Electronics in Agriculture, 218, 108729. https://doi.org/https://doi.org/10.1016/j.compag.2024.108729

Chamberlain, A.T. et al. (2022) 'The relationship between on-farm environmental conditions inside and outside cow sheds during the summer in England: can Temperature Humidity Index be predicted from outside conditions?', Animal - Open Space, 1(1), p. 100019. Available at: https://doi.org/https://doi.org/10.1016/j.anopes.2022.100019.

Cortes, C. and Vapnik, V. (1995) 'Support-vector networks', Machine Learning, 20(3), pp. 273–297. Available at: https://doi.org/10.1007/BF00994018.

Cover, T. and Hart, P. (1967) 'Nearest neighbor pattern classification', IEEE Transactions on Information Theory, 13(1), pp. 21–27. Available at: https://doi.org/10.1109/TIT.1967.1053964.

Cox T.F. and Cox M.A.A. (2000) Multidimensional Scaling. CRC Press. New York.

Dulhare, U.N., Ahmad, K. and Ahmad, K.A. Bin (2020) Machine learning and big data: concepts, algorithms, tools and applications. John Wiley & sons.

Everitt, B. et al. (2011) 'Cluster analysis'.

Frades, I. and Matthiesen, R. (2010) 'Overview on Techniques in Cluster Analysis', in R. Matthiesen (ed.) Bioinformatics Methods in Clinical Research. Totowa, NJ: Humana Press, pp. 81–107. Available at: https://doi.org/10.1007/978-1-60327-194-3_5.

Fuentes, S. et al. (2020) 'Artificial Intelligence Applied to a Robotic Dairy Farm to Model Milk Productivity and Quality based on Cow Data and Daily Environmental Parameters', Sensors 2020, Vol. 20, Page 2975, 20(10), p. 2975. Available at: https://doi.org/10.3390/S20102975.

Giannone, C., Bovo, M., Ceccarelli, M., Torreggiani, D., & Tassinari, P. (2023). Review of the Heat Stress-Induced Responses in Dairy Cattle. Animals, 13(22). https://doi.org/10.3390/ani13223451.

Hinton, G.E. and Roweis, S.T. (2002) 'Stochastic Neighbor Embedding', in Neural Information Processing Systems. Available at: https://api.semanticscholar.org/CorpusID:20240.

Ji, B. et al. (2022) 'A machine learning framework to predict the next month's daily milk yield, milk composition and milking frequency for cows in a robotic dairy farm', Biosystems Engineering, 216, pp. 186–197. Available at: https://doi.org/https://doi.org/10.1016/j.biosystemseng.2022.02.013.

Jia, W. et al. (2022) 'Feature dimensionality reduction: a review', Complex & Intelligent Systems, 8(3), pp. 2663–2693. Available at: https://doi.org/10.1007/s40747-021-00637-x.

Jones, T. (1997) 'Empirical Bayes Prediction of 305-Day Milk Production', Journal of Dairy Science, 80(6), pp. 1060–1075. Available at: https://doi.org/https://doi.org/10.3168/jds.S0022-0302(97)76031-4.

Kino, E. et al. (2019) 'Exploration of factors determining milk production by Holstein cows raised on a dairy farm in a temperate climate area', Tropical Animal Health and Production, 51(3), pp. 529–536. Available at: https://doi.org/10.1007/s11250-018-1720-6.

Lovarelli, D., Bacenetti, J. and Guarino, M. (2020) 'A review on dairy cattle farming: Is precision livestock farming the compromise for an environmental, economic and social sustainable production?', Journal of Cleaner Production, 262, p. 121409. Available at: https://doi.org/https://doi.org/10.1016/j.jclepro.2020.121409.

van der Maaten, L. and Hinton, G. (2008) ‘Visualizing Data using t-SNE’, Journal of Machine Learning Research, 9(86), pp. 2579–2605. Available at: http://jmlr.org/papers/v9/vandermaaten08a.html.

Masía, F.M. et al. (2020) ‘Modeling variability of the lactation curves of cows in automated milking systems’, Journal of Dairy Science, 103(9), pp. 8189–8196. Available at: https://doi.org/https://doi.org/10.3168/jds.2019-17962.

McInnes, L., Healy, J. and Melville, J. (2020a) ‘UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction’.

McInnes, L., Healy, J. and Melville, J. (2020b) ‘UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction’.

Mead, A. (1992) ‘Review of the Development of Multidimensional Scaling Methods’, Journal of the Royal Statistical Society Series D: The Statistician, 41(1), pp. 27–39. Available at: https://doi.org/10.2307/2348634.

Mignotte, M. (2011) ‘MDS-Based Multiresolution Nonlinear Dimensionality Reduction Model for Color Image Segmentation’, IEEE Transactions on Neural Networks, 22(3), pp. 447–460. Available at: https://doi.org/10.1109/TNN.2010.2101614.

Ozella, L. et al. (2023) ‘A Literature Review of Modeling Approaches Applied to Data Collected in Automatic Milking Systems’, Animals, 13(12). Available at: https://doi.org/10.3390/ani13121916.

Perez Garcia, C. A., Bovo, M., Torreggiani, D., Tassinari, P., & Benni, S. (2023). 3D numerical modelling of temperature and humidity index distribution in livestock structures: a cattle-barn case study. Journal of Agricultural Engineering, 54(3). https://doi.org/10.4081/jae.2023.1522.

Perez Garcia, C. A., Bovo, M., Torreggiani, D., Tassinari, P., & Benni, S. (2024). Indoor Temperature Forecasting in Livestock Buildings: A Data-Driven Approach. Agriculture, 14(2). https://doi.org/10.3390/agriculture14020316.

Rebuli, K.B. et al. (2023) ‘Multi-algorithm clustering analysis for characterizing cow productivity on automatic milking systems over lactation periods’, Computers and Electronics in Agriculture, 211, p. 108002. Available at: https://doi.org/10.1016/j.compag.2023.108002.

Strpić, K., Barbaresi, A., Tinti, F., Bovo, M., Benni, S., Torreggiani, D., Macini, P., & Tassinari, P. (2020). Application of ground heat exchangers in cow barns to enhance milk cooling and water heating and storage. Energy and Buildings, 224, 110213. https://doi.org/https://doi.org/10.1016/j.enbuild.2020.110213.

Tenenbaum, J.B., De Silva, V. and Langford, J.C. (2000) ‘A global geometric framework for nonlinear dimensionality reduction’, Science, 290(5500). Available at: https://doi.org/10.1126/science.290.5500.2319.

Tullo, E., Finzi, A. and Guarino, M. (2019) ‘Review: Environmental impact of livestock farming and Precision Livestock Farming as a mitigation strategy’, Science of the Total Environment [Preprint]. Available at: https://doi.org/10.1016/j.scitotenv.2018.10.018.

Velliangiri, S., Alagumuthukrishnan, S. and Thankumar joseph, S.I. (2019) ‘A Review of Dimensionality Reduction Techniques for Efficient Computation’, Procedia Computer Science, 165, pp. 104–111. Available at: https://doi.org/https://doi.org/10.1016/j.procs.2020.01.079.

Witten, I.H., Frank, E. and Hall, M.A. (2011) Data Mining: Practical Machine Learning Tools and Techniques. 3rd edn. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Wood, P.D.P. (1967) ‘Algebraic Model of the Lactation Curve in Cattle’, Nature, 216(5111), pp. 164–165. Available at: https://doi.org/10.1038/216164a0.

Zhou, M. et al. (2022) ‘Effects of increasing air temperature on physiological and productive responses of dairy cows at different relative humidity and air velocity levels.’, Journal of dairy science, 105(2), pp. 1701–1716. Available at: https://doi.org/10.3168/jds.2021-21164.