Application of machine learning methods to control the milk samples analysis results reliability

I.V. Rukin[1], S.I. Sudarkina[1], M.S. Krutkina[1]and E.V. Kamaldinov[2]

[1]JSC «Agroplem», Moscow, Russian Federation
[2]Novosibirsk State Agrarian University, Novosibirsk, Russian Federation
Corresponding Author: irukin@agroplem.ru

Abstract

Modern animal breeding methods, such as genomic evaluation of breeding values (GEBV), are based on large amounts of phenotypic and genetic data. The reliability of GEBV results and the general selection process depends on the accuracy of the primary phenotypic data. Test day milk samples implied to be collected from unique cows may actually be dispensed from a sampler or milk tank. We will further call such milk samples dispensed or DS. To address this issue, a new DS identification system using clustering algorithm (OPTICS) was developed to improve accuracy in detecting DS in milk samples. Results showed high accuracy in identifying DS in small batches, when samples were not dispensed sequentially or were mixed with unique samples. Large batches with more than 60 DS in each were also accurately detected. However, the algorithm showed low accuracy on batches with low DS proportion. This new method has already been implemented in the milk analysis laboratory and will continue to be refined for better data filtering in breeding value systems.

Introduction

GEBV plays the key role in modern methods of livestock production and selection work. Yearly the number of farms including GEBV in their work raises significantly (Song *et al*., 2023) leading to great increase of data collected and analyzed. While GEBV calculations take into account as much available livestock data as possible, the unreliable data may lead to bias and mistakes in GEBV results and erroneous conclusions in selection work. That is why data quality control is crucial process of data preprocessing before GEBV (Cabrera *et al*., 2020).

One of the primary categories of traits in dairy cattle is milk traits, often assessed through TD (test-day) milk samples analyzed in milk laboratories. Research suggests that one potential factor leading to skewed TD milk results is the collection of samples from tanks, rather than individual cows. It's important to identify samples collected from tanks and exclude them from GEBV (genomic estimated breeding value) analysis. While our laboratory acknowledges batches containing dispensed samples collected from tanks (DS) in sequential order, identifying DS samples mixed with unique samples in a batch is more challenging.

The aim of our work is creation of more accurate recognition of DS system. The main objectives of this study are:

- The recognition of DS in the TD samples batch
- Identification of DS for subsequent data filtering

Material and methods

TD milking data collected from 2019 to 2024 was used to generate test datasets. These datasets included batches with varying amounts of dispensed samples (DS), ranging from 0 to 121 DS per batch. We generated a total of three datasets:

- **Dataset 1:** Comprised of 1019 batches, 994 of which contained between 15 to 100 DS each.
- **Dataset 2:** Comprised of 4298 batches, 4289 of which contained between 15 to 100 DS each.
- **Dataset 3:** Comprised of 1000 batches, 997 of which contained between 15 to 121 DS each.

The generation of DS was done as follows:

- **Dataset 1:** In each batch, one sample was chosen randomly. Its fat and protein content were used to generate 15 to 100 points with mean values equal to the fat and protein content of the chosen sample, and a standard deviation of 0.1. These generated points were added to the batch data file in random strings, mixing DS with unique samples. Consequently, the samples in Dataset 1 are generated as if dispensed from one tank.
- **Dataset 2:** The generation method was similar to Dataset 1, but with a variation in the number of samples chosen to generate DS. Here, the number of samples used to generate DS varied randomly from 1 to 10. Thus, Dataset 2 represents batches with DS obtained from multiple tanks.
- **Dataset 3:** This dataset consists of batches with or without DS from one tank, similar to Dataset 1. However, the number of DS per batch varied depending on batch size: 15-61 DS in small batches, 29-101 DS in medium batches, and 59-121 DS in large batches.

A summary of the generated datasets is shown in Table 1.

To improve the quality control algorithm and recognize DS mixed with unique samples in a batch, we applied unsupervised machine learning. We developed an algorithm based on clustering, utilizing the density-based method OPTICS (Ordering Points To Identify the Clustering Structure) (Ankerst *et al*., 1999), available in the Python scikit-learn module (Pedregosa *et al*., 2011). The core idea of the algorithm is to identify clusters of high-density points in the space of milk sample parameters.

We focused on two milk sample parameters obtained from Fossomatic: fat and protein content. The OPTICS clustering algorithm takes the data to be clustered and the

Table 1. Description of generated datasets.

Dataset	Number of batches	Number of batches with DS	Number of tanks in batch	Number of samples in one tank
1	1019	994	0 or 1	15-100
2	4298	4289	0 or 1-10	15-100
3	1000	997	0 or 1	15-61[1] 29-101[2] 59-121[3]

[1]Small size batches, batch size < 150 samples
[2] Medium size batches, 150 < batch size < 800 samples
[3] Large size batches, batch size > 800 samples

clustering parameters: min_samples (the minimum number of points, MinPts) and max_eps (the maximum distance for clustering).

For identifying batches containing DS, we used max_eps = 0.2. The min_samples value varied based on the number of samples in a batch:

- For small batches (fewer than 150 samples), we used min_samples = 15.
- For medium batches (150 to 800 samples), we used min_samples = 30.
- For large batches (more than 800 samples), we used min_samples = 60.

Results and discussion

We tested the clustering algorithm on the three datasets described above and calculated metrics to evaluate the quality of clustering. First, we assessed the algorithm's ability to recognize batches containing DS. According to the algorithm, a batch is considered to contain DS if more than one cluster is found. The calculated performance statistics and metrics are shown in Table 2 and Table 3, respectively.

The metrics used to evaluate the algorithm's performance quality included:

- Rand Index (RI).
- Adjusted Rand Index (ARI).
- Mutual Information (MI).
- Adjusted Mutual Information (AMI).
- V-measure.
- Homogeneity.
- Completeness.

The results shown in Table 2 display fine algorithm performance on small batches. However, the performance on Datasets 1 and 2 decreases with an increase in batch size, a trend not observed in Dataset 3. This performance decline is presumably associated with the proportion of DS in a batch. As the batch size increases, more samples have similar fat and protein content values, making it harder for the algorithm to determine if a small collection of points is DS. With increasing batch size, the min_samples parameter (the number of samples in a neighbourhood for a point to be considered a core point) also increases. As a result, small clusters of DS cannot be properly detected with this method. Not increasing the min_samples parameter with

Table 2. Clustering performance statistics.

		Dataset 1		Dataset 2		Dataset 3	
		True	False	True	False	True	False
All batches	Positive	881	4	4225	1	982	3
	Negative	21	113	4	168	15	0
Small batches	Positive	60	0	305	0	63	1
	Negative	2	1	2	6	3	0
Medium batches	Positive	551	3	2509	1	595	0
	Negative	13	52	2	34	9	0
Large batches	Positive	269	1	1411	0	324	2
	Negative	6	60	0	128	3	0

Table 3. Clustering performance metrics.

		V-measure	RI	ARI	MI	AMI	Homogeneity	Completeness
Dataset 1	All batches	0.13	0.80	0.20	0.03	0.13	0.29	0.08
	Small batches	0.66	0.97	0.78	0.11	0.66	0.78	0.58
	Medium batches	0.16	0.84	0.26	0.04	0.16	0.31	0.11
	Large batches	0.07	0.70	0.10	0.02	0.07	0.22	0.04
Dataset 2	All batches	0.03	0.93	0.04	0.002	0.03	0.27	0.01
	Small batches	0.31	0.96	0.38	0.02	0.30	0.63	0.20
	Medium batches	0.06	0.97	0.10	0.002	0.06	0.29	0.04
	Large batches	0.00	0.85	0.00	0.00	0.00	1.00	0.00
Dataset 3	All batches	0.83	0.99	0.90	0.06	0.83	0.77	0.90
	Small batches	0.73	0.97	0.84	0.15	0.72	0.66	0.82
	Medium batches	1.00	1.00	1.00	0.07	1.00	1.00	1.00
	Large batches	0.64	0.99	0.74	0.04	0.64	0.53	0.80

increasing batch size would lead to a rapid growth in the false positive rate by extracting false, occasional clusters.

Due to the apparent dependence of the algorithm's performance on the proportion of DS in a batch, we decided to generate and analyse a dataset with an increasing number of DS corresponding to the increasing batch size (Dataset 3). The performance of the algorithm on Dataset 3 shows zero false-negative results with a quite low false positive rate, effectively avoiding Type II errors.

Regarding the accuracy of the tests carried out, our algorithm can detect batches with DS if the batch is small or if the DS tank is big enough (more than 30 and 60 samples in medium and large batch respectively). The identification of small number of samples in large batches is still difficult. For further development of the algorithm, we plan to aim our work at:

- Development of method to detect small DS clusters in large batches properly,
- Development of method to choose the proper clustering parameters to detect every serial number of DS properly.
- Development of an algorithm to choose clustering parameters for accurate identification of DS within a batch.

List of references

Ankerst, M., Breunig, M.M., Kriegel, H.-P., Sander, J., 1999. OPTICS: ordering points to identify the clustering structure, in: Proceedings of the 1999 ACM SIGMOD International Conference on Management of Data. Presented at the SIGMOD/PODS99: International Conference on Management of Data and

Symposium on Principles of Database Systems, ACM, Philadelphia Pennsylvania USA, pp. 49–60. https://doi.org/10.1145/304182.304187

Cabrera, V.E., Barrientos-Blanco, J.A., Delgado, H., Fadul-Pacheco, L., 2020. Symposium review: Real-time continuous decision making using big data on dairy farms. J. Dairy Sci. 103, 3856–3866. https://doi.org/10.3168/jds.2019-17145

Pedregosa, F., G. Varoquaux, A. Gramfort, V.M. Bertrand Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, 2011. Scikit-learn: Machine Learning in Python. the Journal of machine Learning research 2825-2830. https://doi.org/10.5555/1953048.2078195

Song, H., Dong, T., Yan, X., Wang, W., Tian, Z., Sun, A., Dong, Y., Zhu, H., Hu, H., 2023. Genomic selection and its research progress in aquaculture breeding. Rev. Aquac. 15, 274–291. https://doi.org/10.1111/raq.12716