Unsupervised Machine Learning for Precision Livestock Welfare

Fei Ge

October 2025

The invited review article explores the advantages of model-free approaches, specifically unsupervised machine learning (UML), over traditional model-dependent statistical methods for analyzing complex data from precision livestock farming (PLF). The authors compare these methodologies, arguing that conventional models can overlook crucial behavioral patterns when upfront assumptions about the management system are incomplete, potentially masking indicators of compromised animal welfare. Through simulated and empirical case studies, including an analysis of milk parlor metadata from an organic dairy, the paper demonstrates how UML, in combination with information-theoretic tools, can effectively recover complex and unanticipated behavioral dynamics and bivariate associations without requiring extensive prior knowledge about the system. The conclusion advocates for using these model-free algorithms as a more open-ended approach to knowledge discovery in the large, noisy datasets common in modern animal science.

1 Introduction

1.1 Simulation Case Study: Data Compression and Information Loss

The goal of this initial simulation was not to categorize the method, but to illustrate a more fundamental concept: how information compression can fail if the assumptions are violated.

1.2 Simulation Case Study: Comparing Model-Dependent and Model-Free Approaches to Information

The authors used a UML method to demonstrate that these approaches can recover complex behavioral patterns even when the environmental factors driving those patterns are not provided to the method.

These sources highlight the limitations of conventional methods and the need for Unsupervised Machine Learning (UML) due to complex PLF data characteristics and the inherent inability of model-dependent approaches to function efficiently without complete prior knowledge of the system.

PLF Data Challenges and Characteristics

PLF technologies allow animal scientists to collect behavioral data on large numbers of animals, over extended time periods, and at high sampling frequencies. This scale and granularity introduce several data complexities that challenge traditional statistical models:

- **Complexity and Nonstationarity:** PLF datasets often contain complex stochastic features that are difficult to accommodate in conventional statistical models. These features include temporal nonstationarity (such as seasonal and circadian rhythms), autocorrelation, heterogeneous variance structures, and nonindependence between experimental units.
- **Remote Data Collection:** Data is frequently collected remotely, meaning complex features or driving environmental factors often cannot be anticipated *a priori*.
- **Data Redundancy:** The sampling frequency in commercially marketed sensor technologies is often dictated by the hardware, leading to a considerable amount of redundancy between data points.
- **Chaos of Commercial Environments:** Moving away from controlled experimental settings toward the chaos and complexity of commercial farm environments makes using model-based approaches for extracting ethological insights fundamentally challenging.

Key Limitations Demonstrated by the Sources

1. **Information Loss and Aggregation Bias**
   A linear model is fundamentally a form of information compression. If the assumptions employed in this compression strategy do not match the reality of the data, biologically relevant patterns can be easily lost.
   - The simulation case study showed that if a model is not correctly structured to capture the system's dynamics, it can lead to aggregation bias, which masks important behavioral indicators of compromised welfare.
   - In the first model-dependent simulation, a consultant using standard EDA (Exploratory Data Analysis) concluded that the herd's average lying time was incredibly stable (stationary), obscuring the pattern. Consequently, the simple linear model found fixed effects for day and heifer status to be insignificant, concluding that there was no evidence of consistent individual differences in lying time. This demonstrated how an inadequate model can become an inefficient means of information compression that "hemorrhages information".
2. **Failure to Capture Nonlinear Dynamics**
   Model-based methods struggle when the links between environmental factors and behavioral responses are strong but nonlinear.
   - The sources describe a simulation where lying time related to the Temperature-Humidity Index (THI) via a threshold model (a nonlinear dynamic).

  - Attempting to find an association using a simple linear effect (Pearson correlation) or even a nonparametric Spearman rank correlation failed to identify THI as a significant influence. The linear model returned a near-zero slope because it was not structured to anticipate this dynamic.

3. **Overwhelming Complexity**
   As complexities in the management system compound, it becomes overwhelming to account for all contingencies within a single model. If the causative mechanisms driving behavioral responses are not simple (e.g., if resource holding potential is determined by numerous unmeasured factors like size, seniority, and aggression), conventional models may fail.

4. **Violation of Statistical Assumptions**
   Complex behaviors like milking order violate nearly every assumption required for drawing statistical inferences from a conventional linear model, namely independence and homogeneity of variance.

Unsupervised Machine Learning (UML) as an Alternative

UML offers a philosophical and technical alternative to overcome the gaps in background knowledge inherent in complex PLF environments.

- **Open-Ended Knowledge Discovery:** UML algorithms are designed to systematically sift through data sets to identify and characterize nonrandom patterns until only noise remains, offering a more flexible and open-ended approach to knowledge discovery. These model-free algorithms require fewer up-front assumptions about the management system.

- **Leveraging Intrinsic Codependencies:** Model-free approaches may overcome gaps in background knowledge by fully leveraging the intrinsic behavioral codependencies of group-housed animals. The simulation demonstrated that Hierarchical Clustering (HC) could successfully recover complex social dynamics (like the inversion of lying patterns) hidden within the data even without being provided information on the causative factors (parity or weather).

- **Handling Nonlinearity:** Model-free information-theoretic approaches, such as those using mutual information estimates, provide a means to identify and characterize complex bivariate associations between datasets regardless of the underlying dynamic (linear, quadratic, exponential, etc.).

- **Enhanced Visualization:** UML tools, particularly clustering algorithms, can be used to extract and visualize the most striking nonrandom features of a data set, making complex patterns visually striking even when obscured in raw data matrices.

Hierarchical Clustering (HC) in Different Scenarios

The Unsupervised Machine Learning (UML) technique used in the section "SIMULATION CASE STUDY: COMPARING MODEL-DEPENDENT AND MODEL-FREE APPROACHES TO INFORMATION COMPRESSION" was Hierarchical Clustering (HC), employed to demonstrate its ability to recover complex social dynamics even when prior knowledge of the causative factors is absent.

The experiment was an extended simulation designed to mimic a real-world scenario involving welfare concerns in an overstocked dairy herd.

Summary of the UML Experiment (Model-Free Approach)

The UML experiment was a specialized version of the larger simulation case study, designed to present a scenario where the causative factors driving behavioral changes were too complex, transient, or unknown to be efficiently captured by model-dependent methods.

Materials (Data and Scenario)

The simulation data analyzed by UML involved a group of 100 overstocked cows. The data stream analyzed was the daily estimates of the proportion of time each animal spent lying down over a 60-day observation interval.

The underlying behavioral patterns were driven by increasing the analytical complexity of the scenario:

1. Overstocking remained the condition driving competitive behavior for limited freestall spaces.

2. Instead of a single, persistent shift in lying patterns (like the start of the grazing season in the model-dependent approach), the inversions in lying patterns were driven by transient environmental factors, specifically randomly scattered rain events.

3. The algorithm was never provided information on the causative factors driving this behavioral pattern (such as cow parity or weather records).

Method (Hierarchical Clustering)

The Hierarchical Clustering (HC) algorithm was used to perform an open-ended approach to information compression.

1. **Dissimilarity Matrix Calculation:** The first step was to compute a dissimilarity matrix. The Euclidean distance (or L2 norm, Equation) was used as the metric to estimate dissimilarity.

    - To cluster cows together with similar lying patterns, the Euclidean distance was calculated between the data vectors for each pair of cows, summed over all observation days.

- To cluster together days with similar lying patterns, the Euclidean distance was calculated between each pair of observation days, summed over all animals in the herd.

2. **Agglomeration:** Ward's (2-dimensional) linkage method was applied as the ground-up agglomeration algorithm. This method merges clusters to produce the largest increase in between-group variance at each step.

3. **Visualization:** The dendrograms produced by clustering cows over days and days over cows were then used to reorder the rows and columns of the raw data matrix. The resulting data matrix was visualized using a heatmap.

Results

The results demonstrated the efficacy of UML in recovering complex dynamics without prior assumptions:

- **Raw Data Obscured Patterns:** When the raw data matrix was visualized without any specific order, the inversions in lying patterns driven by the randomly scattered rain events were completely obscured.

- **Recovery of Social Dynamics:** After reordering the rows (cows) and columns (days) using the HC dendrograms, the visualization became visually striking.

- **Discovery:** The clustering algorithm captured two distinct groups of cows and two distinct groups of observation days, revealing the inversion in lying patterns.

- **Conclusion:** The HC algorithm successfully recovered the social dynamics hiding within the lying time data, thereby indicating that overstocking was compromising the herd's welfare, even though the algorithm was never provided information on the causative factors (parity and weather records). This shows how model-free approaches can overcome gaps in background knowledge by leveraging the intrinsic behavioral codependencies of group-housed animals.

1.3 Simulation Case Study: Comparing Model-Dependent and Model-Free Approaches to Identifying Bivariate Associations

This section details a simulation case study specifically designed to compare model-dependent and model-free approaches in their ability to identify and characterize bivariate associations between a behavioral response and an environmental factor, particularly when the relationship is strong but nonlinear.

Materials (Scenario and Data)

The scenario was a continuation of the previous consulting example, but with management improvements:

- **The Problem:** The farmer had reduced the stocking rate (to a 1:1 ratio), but cows still exhibited inadequate lying times on some days, necessitating further analysis.
- **Data Analyzed:** The proportion of time cows spent lying down was compared against the Temperature-Humidity Index (THI), a candidate environmental variable.
- **Underlying Dynamic (The "Reality"):** The relationship between lying time and THI was simulated as a nonlinear threshold model.
  - At low THI, cows spent the majority of their day grazing (low lying time).
  - As THI gradually rose, cows became heat stressed and increased their time lying down in the shade.
  - Above a certain high THI threshold, cows struggled to thermoregulate when lying down and instead stood for extended periods of time.

This specific dynamic, characterized by a behavioral response subject to competing underlying behavioral response mechanisms, is commonly found in real-world data.

Method and Results: Model-Dependent Approach (Failure)

The sources demonstrate that model-based methods, which assume a particular structure (like linearity), fail to capture this complex dynamic:

- **Method:** A simple linear effect (such as a Pearson correlation test) was used, which assumes a linear association. A nonparametric Spearman rank correlation test was also applied, which assumes a monotonically increasing relationship.
- **Results:**
  - If a simple linear effect was used, a near-zero slope was returned, as the linear model was not structured to anticipate the dynamic.
  - The Pearson correlation test failed to identify the pattern ($r = -0.03$, $P = 0.25$).
  - The nonparametric Spearman rank correlation test also failed ($\rho = 0.03$, $P = 0.12$) because the pattern was not monotonically increasing.
- **Conclusion:** This outcome highlights that model-based approaches can overlook causes of compromised welfare when the link between environmental factors and behavioral responses is strong but nonlinear.

Method and Results: Model-Free Approach (Success)

For applications where intuition about the underlying dynamics is more important than prediction, the information-theoretic framework provides a model-free solution.

- **Method (Information-Theoretic Approach):**

1. **Discretization:** Both the THI and lying time variables were discretized (converted into categorical encodings) using simple equal-sized binning rules (e.g., 4 levels for THI and 7 levels for lying time).
2. **Mutual Information (MI) Test:** Mutual information ($I(X, Y)$) was calculated to reflect the strength of the bivariate association between the two encoded datasets, regardless of the underlying dynamic (linear, quadratic, exponential, etc.). MI quantifies how much information is learned about one variable if the value of the other is known.
3. **Permutation Test:** The observed MI estimate was compared against estimates generated from nonconditional permutations (a standard nonparametric permutation test).
4. **Characterization:** A contingency table was used to visualize the joint distribution, colored by pointwise mutual information (PMI) estimates, which show how much each joint encoding differs from what would be expected if no association existed.

- **Results:**
  - The nonparametric test using information entropy as the test statistic found a highly significant bivariate association ($P \leq 0.001$).
  - The PMI visualization clearly demonstrated that the probability of observing a given lying pattern is shifted in different directions based on the level of the THI encoding. For example, orange cells indicated overrepresentation of specific joint counts, while blue cells indicated underrepresentation.
- **Conclusion:** This model-free approach successfully identified a significant bivariate association and provided insights into the underlying dynamic to inform management interventions, absent any prior assumptions about the relationship between the two variables.

1.4 Empirical Case Study: Model-Free Knowledge Discovery with Milk Parlor Metadata

This empirical case study utilized a model-free analytical pipeline to uncover complex, unanticipated behavioral patterns hidden within milk parlor metadata from a commercial organic dairy farm. The goal was to demonstrate how Unsupervised Machine Learning (UML) and information-theoretic tools can be seamlessly integrated to analyze data with minimal assumptions.

Materials (Setting and Data)

The data, known as the Organilac data set, was collected in 2017 from a USDA-certified organic dairy in northern Colorado.

- **Herd Structure:** The study involved 200 cows enrolled over 1.5 months into a mixed-parity herd, primarily of Holstein genetics.
- **Housing and Management:** Cows were kept in an open-sided freestall barn stocked at roughly half capacity, with free access to TMR and an adjacent outdoor dry lot. Starting in April, cows also had free access to pasture at night.
- **Data Source:** Cows were milked three times a day in an RFID-equipped rotary parlor.
- **Key Variables:** The parlor recorded daily milk production weights and milking order (the sequence in which cows entered the parlor).

The study specifically focused on analyzing the complex associations between milking order, cow age structure, and levels of milk production. Conventional analysis of milking order is challenging because the behavior violates nearly every assumption required for drawing statistical inferences from a conventional linear model, such as independence and homogeneity of variance.

Methods

The analysis utilized the Livestock Informatics Toolkit (LIT), an open-source software package.

1. Unsupervised Encoding (Hierarchical Clustering) Hierarchical Clustering (HC) was used to encode the continuous data for age and yield:

- **Data Used:** Cow age in days and the 95th daily yield quantile (used as a model-free estimate of peak yield).
- **Rationale:** Age and yield are often highly correlated; clustering was used to leverage this redundancy to establish empirically determined cutoffs.
- **Encoding Results:** Examination of the dendrogram used to order the rows revealed that the clustering reflected the redundancy between age and yield, isolating heifers first, then distinguishing still-growing second-lactation cows from fully grown cows (parity 4 or more), with third-lactation animals divided based on yield.

2. Model-Free Association Testing (Bivariate Tree Test) A model-free bivariate tree testing framework was applied using information entropy/mutual information (MI) to test for associations between the age/yield encoding and previously reported encodings of queuing patterns.

- **Rationale:** The tree test performs MI analysis across all combinations of encoding granularities (cluster numbers) for both data sets to avoid losing valuable information due to using an arbitrary cluster count.

- **Result (Optimal Association):** A highly significant bivariate association ($P = 0.001$) was identified when using 3 clusters for the age/yield encoding and 6 clusters for the queuing pattern encoding. The strongest association was consistently found when the entry order data was encoded using 2 temporal subperiods to capture non-stationarity.

3. Characterization (Pointwise Mutual Information – PMI) A contingency table was used to characterize the dynamics of the significant association, with cells colored by their pointwise mutual information (PMI) estimates. This visualization highlights encoding combinations that are significantly overrepresented (orange) or underrepresented (blue) compared to the expectation if no association existed.

Results

The model-free analysis revealed a clear, inverted age dynamic governing milking order.

1. Inverted Age Dynamic at Front vs. Rear of Queue The analysis found that the relationship between age/yield and queue position was nonlinear and complex:

- **Front of Queue (Queue Cluster No. 4):** The oldest cows in the herd (age yield cluster no. 2) were significantly overrepresented among animals that consistently entered at the very front of the queue.
- **Rear of Queue (Queue Cluster No. 1):** The oldest cows were significantly underrepresented, while the second- and third-lactation animals (age yield cluster no. 4) were significantly overrepresented.

The interpretation offered for this pattern is that the largest and most experienced animals lead the queue, while the second- and third-lactation animals may fill a “caboose cow” role, driving smaller or less fit animals forward.

2. Complex Temporal Shifts The model-free approach also captured a more temporally complex queuing pattern related to pasture access:

- **Queue Cluster No. 5:** Described animals that entered nearer the front or middle of the queue when arriving from the home pen, but fell back nearer the rear when returning to the parlor from overnight pasture.
- The oldest cows (age yield cluster no. 3) were significantly overrepresented in this cluster.

This finding suggests that these older cows might be experiencing compromise in fitness or welfare during the pasture subperiod, which causes them to be pushed backward in the queue by younger herdmates. The authors note that this inverted age dynamic would have been challenging to capture with a model-based analysis using simple linear terms.