

# Association Rule Mining to Study Process-Related Cause-Effect-Relationships in Pig Farming

Tobias Zimpel<sup>1</sup>, Andrea Wild<sup>2</sup>, Hansjörg Schrade<sup>2</sup> and Stefan Kirn<sup>1</sup>

<sup>1</sup>University of Hohenheim, Schloß Hohenheim 1, Stuttgart, 70599, Germany

<sup>2</sup>Boxberg Teaching and Research Centre, Seehöfer Str. 50, Boxberg, 97944, Germany

## Abstract

Association rule mining is a technique for discovering relationships in large data sets and thus can obtain insights into process-related phenomena described by the data. In pig farming, necrosis (dead tissue) in the rearing period is an important phenomenon because it negatively affects animal welfare and reduces the share of usable young pigs. Pig rearing is a long-lasting process involving multiple stakeholders and management activities. Causes of necrosis are often unknown and their identification requires considerable time and effort. Pig rearing is a long-lasting process involving multiple stakeholders and management activities. Causes of necrosis are often unknown and their identification requires considerable time and effort. The objectives of this research are to (1) develop an association rule mining approach for generating plausible suggestions for cause-effect relationships of necrosis in pig rearing and (2) empirically evaluate the predictive power of the discovered rule set. We propose a procedure for generating and comparing association rules for different aggregation intervals. We used data from 672 pigs, collected over a ten-month period (Oct. 2018 - Apr. 2019 & Oct. 2019 - Dec. 2019). Association rules were created on the training set and tested on the test set. Association rules were evaluated using the metrics support, confidence, and lift, and expert knowledge. The association rules focused on temperature-related attributes and achieved confidence values between 0.65 and 0.99. Association rules suggest temperature and underlying processes as (contributing) causes of necrosis. Process experts provided knowledge that supports these suggestions and indicates their plausibility.

## Keywords

association rule mining, pig farming, cause-effect-relationships, process-related data analysis

## 1. Introduction

Association rule mining is a popular technique for creating relationships between attributes in data sets. Therefore, association rule analysis can analyze process-related from different sources to obtain insights into relevant process phenomena [1, 2]. In pig farming, tail necrosis (dead tissue) is such an phenomena in the rearing period [3].

Pig rearing is often a seven-week process for increasing pigs' weight (e.g., from 5 to 25 kg), that involves multiple sub-processes, such as providing food or controlling the ventilation and heating system in a pen (pigs' environment). These sub-processes affect pigs and their environment (e.g., air temperature) and can trigger conditions for necrosis development (e.g., tail biting due to stress) [4, 5, 6]. While sensors and corresponding data processing are present


---

PMAI@IJCAI22: International IJCAI Workshop on Process Management in the AI era, July 23, 2022, Vienna, Austria

✉ tobias.zimpel@uni-hohenheim.de (T. Zimpel); andrea.wild@lsz.bwl.de (A. Wild); hansjoerg.schrade@lsz.bwl.de (H. Schrade); stefan.kirn@uni-hohenheim.de (S. Kirn)



© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

(e.g., to detect pigs [7] or predict losses [8]), the development of necrosis and its causes are usually unobserved or unknown (e.g., due to fewer specific sensors concerning the process output or the presence of non-speaking stakeholders). We assume that the causes of necrosis lie in the process. Therefore, association rule mining may suggest cause-effect relationships.

Previous work propose suggestions for causes should be based on association rules for a training set and a test set (as commonly used for supervised machine learning tasks), differentially preprocessed process data, and the incorporation of process-related knowledge [2, 9, 10, 11, 12]. However, association rules describe a correlation and not a causality between attributes. Thus, we refer to the concept of interestingness to assess possible suggestions. Interestingness depends on the application and is objective or subjective [13, 14]. Objective interestingness is calculated based on the underlying data (e.g., confidence), while subjective interestingness is given when the association rule is unexpected or actionable for the user [15, 14]. However, metrics and the properties unexpected and actionable do not include process knowledge.

We propose the plausibility property to assess suggestions. We call a suggestion plausible if it is justified by process knowledge and based on underlying objective interesting association rules. Justification by process knowledge can be integrated into the analysis of the created association rules [16]. Objective interestingness is measured in this work using the metrics of support, confidence and lift. While lift is calculated on support, support and confidence are already necessary metrics for algorithm configuration with respect to minimum support and minimum confidence [17, 18]. Minimum support and minimum confidence influence the creation of association rules in terms of inclusion or exclusion of attributes and rules, while the algorithm decides how association rules are created [17, 18]. Consequently, association rules and their analysis depend mainly on the data preprocessing, e.g., the discretization of the continuous values. Against this backdrop, we address the following research question: *How to design a procedure to create plausible association rules for suggestions of cause-effect-relationships in pig rearing?* The main contributions of this paper are as follows:

- We investigate association rule mining in terms plausible suggestions of cause-effect relationships using the example of pig farming.
- We propose a procedure for creating association rules based on data from three aggregation intervals (hourly, daily, weekly) to serve as the basis for proposing causes.
- We evaluate generated association rules by using an independent test set and justifications based on the process knowledge.

Our paper is structured as follows: In section 2, we provide background literature. In section 3, we describe the data set and procedure. In section 4, we analyze generated association rules by integrating process-knowledge, and in section 5, we provide a brief conclusion.

## 2. Background

### 2.1. Association rule mining

Association rule mining describes an approach to describing relationships in data by processing binary coded data to create association rules [19]. Given is a set  $S = \{s_0, \dots, s_n\}$  with  $n \in \mathbb{N}$  elements  $s$  (e.g.,  $s$  is a pig). Each element  $s$ , in turn, is a set with up to  $m \in \mathbb{N}$  attributes of the

set  $A$ .  $A$  contains all possible attributes of  $s$ . Thereby,  $s$  has an attribute  $a \in A$ , if  $a$  is true for  $s$  (e.g.,  $\{necrosis\} \in s$  if a pig has necrosis). Given is a set  $X \subset A$ , a set  $Y \subset A$ , and  $X \cap Y = \emptyset$  we define an association rule as a relationship between the two sets as follows [19]:

$$X \Rightarrow Y \quad (1)$$

Relationship means,  $Y$  is probably true if  $X$  is given [20]. To assess this relationship, we use the metrics of *support* ( $supp$ ), *confidence* ( $conf$ ), and *lift*. *Supp* describes how frequent a subset of attributes is in  $S$ . *Conf* describes, how often an association rule is true. *Lift* describes if  $X \cup Y$  occur less or more often than expected. These metrics are calculated as follows [19, 21]:

$$supp(X) = \frac{|\{s \in S | X \subseteq s\}|}{|S|} \quad (2)$$

$$conf(X \Rightarrow Y) = \frac{supp(X)}{supp(X \cup Y)} \quad (3)$$

$$lift(X \Rightarrow Y) = \frac{supp(X \cup Y)}{supp(X) \cdot supp(Y)} \quad (4)$$

$Lift(X \Rightarrow Y) = 1$  means  $X$  and  $Y$  are independent, while  $lift(X \Rightarrow Y) > 1$  indicates a positive correlation and  $lift(X \Rightarrow Y) < 1$  indicates a negative correlation. The process of association rule generation consists of two steps. First: generation of a candidate set  $CS$  consisting of several sets  $X$  and or  $Y$  with *minimal support* ( $supp_{min}$ ). Second, deriving association rules based on  $CS$  with *minimal confidence* ( $conf_{min}$ ).  $Supp_{min}$  affects  $|CS|$  and thus the theoretically number of association rules as well. The selection of  $supp_{min}$  can be set manually for all attributes or for each individual attribute [22, 23]. In practice, however, there may be difficulties in determining  $supp_{min}$  because of the tradeoff between the number of association rules that can be processed by humans and the loss of rules (and thus of suggestions for causes) [2].

## 2.2. Association rule mining for process-related cause-effect relationships

[2] propose association rule mining for analyzing problems in a drill manufacturing process. The report on the need to include process experts for the subsequent analysis of the association rules in order to improve manufacturing processes. [9] analyzed construction project accidents by comparing association rules for two types of projects. The association rules created were applicable to only one of the two project types with one exception (with varying confidence), indicating the specificity of the association rules with respect to the training data. [11] also analyzed accidents on construction sites. They report different confidence values for the same rules on different training sets, suggesting that training sets with different focus support the analysis of association rules. They also show that process-related knowledge can support the analysis by containing information that is not in the data set. [12] investigated faults of distribution terminals in power supply networks. They report on a reduced downtime by performing maintenance based on association rules. While this suggests that association rule mining could find suggestions for causes, it is unclear whether the association rules describe actual causes or whether the reduced downtime is a result of more intensive (e.g., earlier)

maintenance. [10] studied diseases in two broiler farms. They report on specific association rules each applicable to only one of two farms. For more general association rules, discretizing the attributes differently (so that an association rule for one farm applies to another farm) and evaluating on independent data for each farm can increase the explanatory power. In summary, a method for plausible association rules consists of multiple preprocessing approaches (in terms of generalization), training and testing, and the inclusion of process knowledge.

### 3. Materials and methods

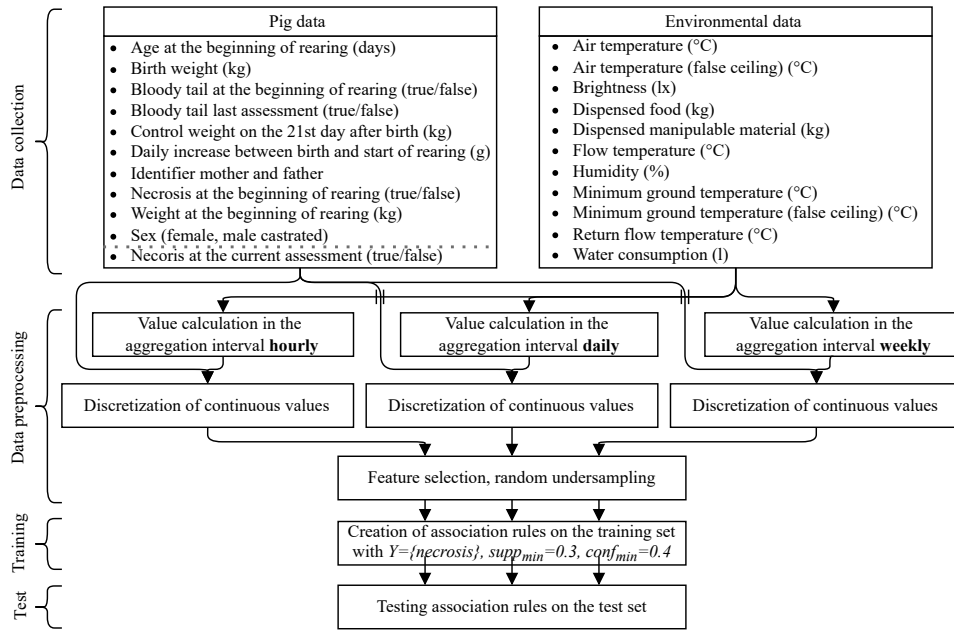
We used process data from pig rearing provided by the Boxberg Teaching and Research Centre in Germany. The Boxberg Teaching and Research Centre is the central educational, experimental, and testing facility of the state of Baden-Württemberg in the field of pig farming. This data set describes the rearing process in the winter season between 10/18/2018 and 4/10/2019 and 10/10/2019 and 12/01/2019. A total of 672 pigs were reared in seven groups of 96 pigs each. Our data set consists of eleven pig-related and eleven environmental factors (figure 1). From our point of view, these factors represent events or results of the underlying processes.

Pig-related factors were weights, sex, daily gain between birth and start of rearing, mother, father, age at the start of rearing, and the presence of necrosis and bloody tails. Weights, sex, daily gain, age at start of rearing, and identification of mother and father were recorded manually. The presence of pig necrosis and bloody tails was recorded during a weekly assessment according to a uniform grading scale (see [24]). Environmental factors were temperatures, humidity, brightness, and water consumption, recorded at five-second intervals, and daily dispensed food and manipulable material. One part of the pen was covered by a false ceiling about 0.75 m high, while the larger part had a ceiling height of 2.9 meters. Thus, air and ground temperatures are divided into temperatures under and outside the false ceiling. Our approach consists of analyzing association rules created for three aggregation intervals of environmental factors: hourly, daily, and weekly (figure 1) [25]. Aggregation intervals means that attributes are calculated on the basis of hourly, daily or weekly periods (e.g., mean humidity in an hour). The implementation uses python 3.8, pandas 1.1.5, scikit-learn 0.24.2, and mlexend 0.19.

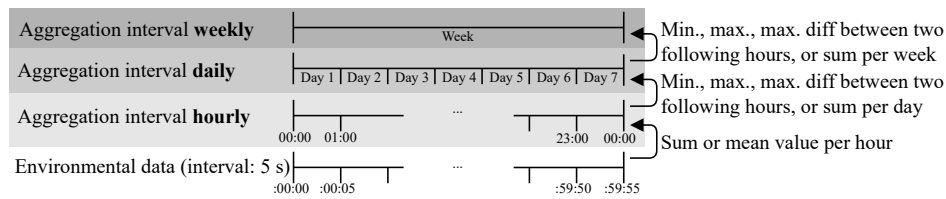
#### 3.1. Data preprocessing

During the preprocessing, we performed the steps: (1) unification of labels and data formats, (2) refilling environmental data, (3) removal of outliers, (4) creation of samples, (5) one hot encoding of categorical features, and (6) creation of training and test sets. The source data of temperatures, water, humidity, and brightness consisted of entries when a change was measured. Missing values were filled in a five-second interval by forward fill. We detected outliers using lower and upper limits set by experts. Outliers (e.g.,  $temperature = 3,000^{\circ}C$ ) are errors and have been removed [26]. In step (4) we combined pig with environmental data for each aggregation interval, resulting in the same environmental attributes for each pig in a pen (figure 2).

For the aggregation interval hourly, we calculated average values for temperatures, humidity, brightness and water for each hour in a week. We also calculated the difference between the flow and return flow temperatures, and the designated amount for food and manipulable materials on a daily basis. For the daily aggregation interval, we used the hourly data to calculate the



**Figure 1:** Procedure for studying cause-effect relationship

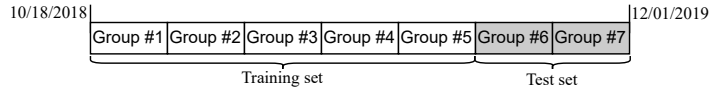


**Figure 2:** Aggregation of environmental data

minimum, maximum, and maximum difference (between two hours) for temperatures, brightness (without minimum), humidity, and the difference between flow and return flow temperatures for each day in a week. We also added the sum of water, food and manipulable material for each day. For the weekly aggregation interval, we calculated the minimum, maximum, and maximum difference between two hours for temperatures, brightness (without minimum), humidity, and the difference between flow and return flow temperatures in a week. We also added the weekly sum and maximum daily difference of water, food, and manipulable material.

Then we assigned continuous values to the corresponding classes with the value 1 (true) if the continuous value is in the corresponding range, and 0 (false) otherwise (discretization into binary values, table A) [27]. Classes were determined in discussion with domain experts.

In step (5), one-hot coding was performed for sex and mother and father identifiers. In step (6), we created our training and independent testing set based on the chronological start of rearing of the groups used (see figure 3). Due to discretization, our training set consists of more attributes than samples, which can lead to overfitting and potentially exponential growth of possible association rules [17, 28]. Therefore, we used standard recursive feature elimination



**Figure 3:** Splitting the data into a training set and a test set

**Table 1**

Overview of the training set and the test set

	Sample size	Quota necrosis	Quota no necrosis
Training set	380	0.50	0.50
Test set	116	0.65	0.35

with random forest as estimator to select 150 features, reducing 50 per round. We also performed random undersampling to balance necrosis in the training set to specify a higher  $supp_{min}$ . A brief description of the training and test set is provided in table 1.

### 3.2. Training

We used the FP-growth algorithm to create association rules because FP-growth does not require candidate generation [29]. The FP-growth algorithm was configured with  $supp_{min} = 0.3$  and  $conf_{min} = 0.4$ . The original intention was to allow rare attributes in association rules ( $supp_{min} = 0.1$ ) since rare traits can also be attributes. However,  $supp_{min} = 0.1$  was not computable in the main memory, so we increased  $supp_{min}$  by  $supp_{min} = 0.1$  until it was computable, resulting in  $supp_{min} = 0.4$ . We chose  $conf_{min} = 0.4$  because we expected a lower value than 0.5 due to unclear multifactorial causes and still wanted to maintain sufficient confidence for practice. We removed all associations rules with  $Y \neq \{necrosis\}$ . We also removed association rules with food-related attributes in  $X$  after discussions suggesting that altered eating behavior may be a consequence of necrosis and thus not a cause.

### 3.3. Testing

For testing, the created association rules were re-identified on the test set using the FP-Growth algorithm and  $supp_{min} = 0.1$  and  $conf_{min} = 0.0000001$ . We re-identified these associations rules to discard rare associations rules in terms of support so that an association rule that only applies to one or two pigs is ignored. Association rules with the highest confidence on the training set tested on the test set are shown in tables 2 (aggregation level hourly), 3 (level daily), and 4 (level weekly).  $t - i$  corresponds to the number of days before an assessment.

Association rules in the hourly aggregation interval on the training set (table 2) focused on the flow temperature and air temperature based on the confidence level and frequency when the time period is not considered. Each association rule consisted of only one combination of the attributes air temperature  $18^{\circ}C \leq x < 21^{\circ}C$ , flow temperature  $20^{\circ}C \leq x < 25^{\circ}C$ , or diff. between flow and return flow temperature  $0^{\circ}C \leq x < 3^{\circ}C$ . The highest confidence and lift on the test set is achieved by rule no. 1, while rule no. 3., which consists of one more temperature attribute in  $X$ , achieves the highest confidence on the training set. According to rule no. 3,

**Table 2**Association rules (aggregation interval hourly,  $Y = \{necrosis\}$ ) on the test set and (training set).

#	Attribute	X			supp	X $\Rightarrow$ Y	
		Period	Function	Range (in °C)		conf	lift
1	Air temperature Diff. flow and return flow tem- perature	t-4 [9h, 10h[ t-3 [9h, 10h[	Mean Mean	$18 \leq x < 21$ $0 \leq x < 3$	0.34 (0.38)	0.98 (0.79)	1.51 (1.57)
2	Air temperature	t-4 [9h, 10h[	Mean	$18 \leq x < 21$	0.34 (0.39)	0.65 (0.70)	1.00 (1.40)
3	Diff. flow and return flow tem- perature	t-3 [9h, 10h[	Mean	$0 \leq x < 3$	0.65 (0.48)	0.98 (0.68)	1.51 (1.36)

the temperature between flow and return flow temperature was  $< 3^\circ C$ , indicating the heating system emitted no or less heat. As a first conclusion, temperature (especially air temperature and flow temperature) and the corresponding heating system are candidates for suggestions on the causes of necrosis. These association rules are assigned to a specific hourly period. To support the first conclusion, we compare this preliminary conclusion with association rules of the daily and weekly aggregation interval.

Association rules at the aggregation interval daily on the training set focused on temperature attributes (see table 3). In particular, maximum air temperature  $18^\circ C \leq x < 21^\circ C$ , and maximum air temperature under the false ceiling  $24^\circ C \leq x < 27^\circ C$ , and return flow temperature  $25^\circ C \leq x < 30^\circ C$  were frequent attributes. Rules no. 4 to no. 6 with the highest confidence level in the training set were not applicable to the test set. The attributes maximum air temperature  $18^\circ C \leq x < 21^\circ C$ , minimum and maximum air temperature under the false ceiling  $24^\circ C \leq x < 27^\circ C$  were also included in other rules during training in other periods, but reached a lower *conf*. No other process environment attributes such as brightness were present in the association rules of the training set. Although the rules in table 3 were not applicable to the test set and the confidence of the association rules in the test set was lower than in the training set, we conclude that the association rules on the aggregation interval daily support the preliminary conclusion that temperature is a possible cause.

At the aggregation interval weekly, the association rules (table 4) on the training set focused on the air temperature under the false ceiling, expressed by two association rules describing the minimum  $21^\circ C \leq x < 24^\circ C$  and maximum  $24^\circ C \leq x < 27^\circ C$  air temperature under the false ceiling. However, the rule no was not applicable on the test set. The air temperature itself occurs only in one association rule, which refers to the maximum difference between two consecutive hours ( $1^\circ C \leq x < 2^\circ C$ ), indicating that the air temperature was almost constant during one week. While another association rule consisted of the maximum difference in manipulable material output between two days, there are no other association rules that do not consist of temperatures. Also based on the association rules for this aggregation interval, we find that temperature is a possible suggestion for the cause for necrosis.

**Table 3**Association rules (aggregation interval daily,  $Y = \{necrosis\}$ ) on the test set and (training set).

#	Attribute	X			X $\Rightarrow$ Y		
		Period	Function	Range (in °C)	supp	conf	lift
4	Return flow temperature	t-3	Max(Mean)	$25 \leq x < 30$	<0.1 (0.32)	n.a. (0.79)	n.a. (1.59)
	Air temperature (false ceiling)	t-6	Max(Mean)	$24 \leq x < 27$			
5	Air temperature (false ceiling)	t-3	Max(Mean)	$24 \leq x < 27$	<0.1 (0.32)	n.a. (0.79)	n.a. (1.59)
	Air temperature (false ceiling)	t-6	Max(Mean)	$24 \leq x < 27$			
	Air temperature	t-7	Max(Mean)	$18 \leq x < 21$			
6	Air temperature (false ceiling)	t-0	Min(Mean)	$24 \leq x < 27$	<0.1 (0.30)	n.a. (0.78)	n.a. (1.55)
	Air temperature (false ceiling)	t-6	Max(Mean)	$24 \leq x < 27$			
	Air temperature	t-7	Max(Mean)	$18 \leq x < 21$			

**Table 4**Association rules (aggregation interval weekly,  $Y = \{necrosis\}$ ) on the test set and (training set).

#	Attribute	X		X $\Rightarrow$ Y		
		Function	Range (in °C)	supp	conf	lift
7	Air temperature (false ceiling)	Min(Mean)	$21 \leq x < 24$	0.59 (0.32)	0.99 (0.67)	1.52 (1.33)
8	Air temperature (false ceiling)	Max(Mean)	$24 \leq x < 27$	<0.1 (0.32)	n.a. (0.64)	n.a. (1.27)
9	Air temperature	Max(Diff. between two hours)	$1 \leq x < 2$	0.34 (0.36)	0.98 (0.61)	1.52 (1.23)

## 4. Results and discussion

Analysis of association rules per aggregation interval in both the training and test propose temperatures as a cause of necrosis in the pig's environment. According to process experts, an air temperature of  $18^{\circ}C < 21^{\circ}C$  indicates a low air temperature. During the study period, tests were conducted in the pens with different heating zones. Air temperatures in the areas under the false ceiling were warmer than in the areas without the false ceiling. Therefore, process-related information is available to support temperatures as a suggestion for the causes of necrosis. Process experts said in discussions that the temperatures in the association rules are quite plausible. Plausibility refers to subjective interestingness by including of process-related knowledge [16, 14]. Thus, we conclude that association rule mining is able to create plausible association rules for use cases where there are few or no possibilities to identify



causes of undesirable process outputs. Lack of specific sensors (e.g., necrosis-specific sensors), non-speaking process stakeholders (e.g., pigs), or an environment that limits reproducibility (e.g., different numbers of pigs affected by necrosis living in a similar environment) are examples of such reasons that make the root cause analysis difficult. Other possible applications include diseases, quality variations and yield fluctuations in agriculture or animal husbandry. To use our procedure in these use cases, the application-specific class adaptation in table A is necessary.

Temperature is associated with underlying processes, like heating or ventilation control. Thus, process-related causes could lie in the underlying processes. Although association rules establish a relationship between temperatures (or underlying processes) and necrosis, they describe no causality. The suggestion that temperature is a cause needs to be investigated in further experiments, e.g., similar to [12].

As in [9, 10, 11], created association rules are partially specific to the training set, resulting in inapplicable association rules on the test set. This can be a result of our discretization scheme and different properties of the test set. Future work relate to the creation of discretization schemes and the transfer of generalization concepts from attributes in hierarchical structures to non-hierarchical structures, as is the case with the environmental attributes [25]. Generalization improves transferability to other use cases. Similar to [9, 11], we report that process-related knowledge improves the downstream analysis of association rules with respect to justification.

This work is subject to the following limitations. First, pig farming consists of sub-processes that are carried out in an orderly structure. Association rules use sets of attributes and therefore do not consider the order of characteristics of the underlying processes. To maintain the ordered structure, sequential rule discovery is one approach [30]. Comparing results using association rule mining or sequential rule mining to study cause-effect relationships is future work. Second, class lower and upper bounds affect the support of concerning attributes, may exclude attributes due to low support. In addition, recursive feature elimination excludes attributes once again. Feature elimination was necessary to ensure computability due to available main memory. Third, we did not remove redundant association rules in the downstream analysis. However, this would not change the association rules presented or the suggestions, but would reduce the number of rules. Fourth, we randomly balanced the samples with and without necrosis in the training set to increase  $supp_{min}$ . This directly affects our metrics in the training set. Fifth, we re-identified the association rules in the test set to treat rare association rules as inapplicable association rules because these rules would affect too few pigs to represent a cause-effect relationship. Thus, it is not possible to distinguish between association rules with  $supp = 0$  or  $0 < supp < 0.1$ .

## 5. Conclusion

This research explored association rule mining for plausible suggestions of process-related cause-effect relationships in pig farming. We used real data over a ten-month period and created association rules based on different intervals of data aggregation. Association rules pointing to temperatures and thus the underlying process regarding heat or air control as a suggestion for causes of necrosis. Moreover, there is process-related knowledge (e.g., different heating zones) that supports this suggestion, so we assume that association rule mining is able to provide plausible (justified by process knowledge) suggestions.

## Acknowledgments

We thank the LABEL-FIT project team, especially Eva Gallmann, William Gordillo, Barbara Keßler, and Svenja Opderbeck for data collection and data provision. This work was supported by the project “Landwirtschaft 4.0: Info-System (Phase 2)”, funded by the Ministry for Food, Rural Areas and Consumer Protection of Baden-Württemberg, Germany. This work was also supported by the project LABEL-FIT by funds of the Federal Ministry of Food and Agriculture (BMEL) based on a decision of the Parliament of the Federal Republic of Germany via the Federal Office for Agriculture and Food (BLE) under the innovation support programme (2819200415).

## References

- [1] S. Schönig, A. Rogge-Solti, C. Cabanillas, S. Jablonski, J. Mendling, Efficient and customisable declarative process mining with sql, in: S. Nurcan, P. Soffer, M. Bajec, J. Eder (Eds.), *Advanced Information Systems Engineering*, volume 9694 of *Lecture Notes in Computer Science*, Springer International Publishing, Cham, 2016, pp. 290–305. doi:10.1007/978-3-319-39696-5\_18.
- [2] S. Nurcan, P. Soffer, M. Bajec, J. Eder (Eds.), *Advanced information systems engineering*, *Lecture Notes in Computer Science*, Springer International Publishing, Cham, 2016. doi:10.1007/978-3-319-39696-5.
- [3] G. Reiner, J. Kühling, M. Lechner, H. Schrade, J. Saltzmann, C. Muelling, S. Dänicke, F. Loewenstein, Swine inflammation and necrosis syndrome is influenced by husbandry and quality of sow in suckling piglets, weaners and fattening pigs, *Porcine health management* 6 (2020) 32. doi:10.1186/s40813-020-00170-2.
- [4] K. I. Fiesjå, I. Solberg, Pathological lesions in swine at slaughter: Iv. pathological lesions in relation to rearing system and herd size, *Acta Veterinaria Scandinavica* 22 (1981) 272–282. doi:10.1186/bf03547516.
- [5] F. Kjell I., I. B. Forus, I. Solberg, Pathological lesions in swine at slaughter: V. pathological lesions in relation to some environmental factors in the herds, *Acta Veterinaria Scandinavica* 23 (1982) 169–183.
- [6] K. I. Fiesjå, H. O. Ulvesæter, Pathological lesions in swine at slaughter: I. baconers, *Acta Veterinaria Scandinavica* 20 (1979).
- [7] M. Riekert, A. Klein, F. Adrion, C. Hoffmann, E. Gallmann, Automatically detecting pig position and posture by 2d camera imaging and deep learning, *Computers and Electronics in Agriculture* 174 (2020) 105391. doi:10.1016/j.compag.2020.105391.
- [8] T. Zimpel, M. Riekert, A. Klein, C. Hoffmann, Machine learning for predicting animal welfare risks in pig farming, *Agricultural Engineering* 76 (2021) 24–35. doi:10.1515/1t.2021.3261.
- [9] C.-W. Cheng, C.-C. Lin, S.-S. Leu, Use of association rules to explore cause–effect relationships in occupational accidents in the taiwan construction industry, *Safety Science* 48 (2010) 436–444. doi:10.1016/j.ssci.2009.12.005.
- [10] S. Maneewongvatana, S. Maneewongvatana, T. Lojitamnuay, M. Juthasong, Using association rules to identify root causes of crd in broilers, in: *The 2013 10th International*

- Joint Conference on Computer Science and Software Engineering (JCSSE), IEEE, 2013, pp. 206–210. doi:10.1109/JCSSE.2013.6567346.
- [11] B. U. Ayhan, N. B. Doğan, O. B. Tokdemir, An association rule mining model for the assessment of the correlations between the attributes of severe accidents, *Journal of Civil Engineering and Management* 26 (2020) 315–330. doi:10.3846/jcem.2020.12316.
- [12] X. Zhang, Y. Tang, Q. Liu, G. Liu, X. Ning, J. Chen, A fault analysis method based on association rule mining for distribution terminal unit, *Applied Sciences* 11 (2021) 5221. doi:10.3390/app11115221.
- [13] R. Agrawal, R. Srikant, Fast algorithms for mining association rules in large databases, in: *Proceedings of the 20th International Conference on Very Large Data Bases, VLDB '94*, Morgan Kaufmann Publishers Inc, San Francisco, CA, USA, 1994, pp. 487–499.
- [14] A. Silberschatz, A. Tuzhilin, On subjective measures of interestingness in knowledge discovery, in: *Proceedings of the First International Conference on Knowledge Discovery and Data Mining, KDD'95*, AAAI Press, 1995, pp. 275–281.
- [15] G. Dong, J. Li, Interestingness of discovered association rules in terms of neighborhood-based unexpectedness, in: X. Wu, R. Kotagiri, K. B. Korb (Eds.), *Research and Development in Knowledge Discovery and Data Mining*, volume 1394 of *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, Berlin, Heidelberg, 1998, pp. 72–86. doi:10.1007/3-540-64383-4\_7.
- [16] G. Piatetsky-Shapiro, C. J. Matheus, The interestingness of deviations, in: *Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining, AAAIWS'94*, AAAI Press, 1994, pp. 25–36.
- [17] Z. Zheng, R. Kohavi, L. Mason, Real world performance of association rule algorithms, in: F. Provost, R. Srikant, M. Schkolnick, D. Lee (Eds.), *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '01*, ACM Press, New York, New York, USA, 2001, pp. 401–406. doi:10.1145/502512.502572.
- [18] J. Hipp, U. Güntzer, G. Nakhaeizadeh, Algorithms for association rule mining – a general survey and comparison, *ACM SIGKDD Explorations Newsletter* 2 (2000) 58–64. doi:10.1145/360402.360421.
- [19] R. Agrawal, T. Imieliński, A. Swami, Mining association rules between sets of items in large databases, *ACM SIGMOD Record* 22 (1993) 207–216. doi:10.1145/170036.170072.
- [20] P. Hájek, I. Havel, M. Chytil, The guha method of automatic hypotheses determination, *Computing* 1 (1966) 293–308. doi:10.1007/bf02345483.
- [21] S. Brin, R. Motwani, J. D. Ullman, S. Tsur, Dynamic itemset counting and implication rules for market basket data, *ACM SIGMOD Record* 26 (1997) 255–264. doi:10.1145/253262.253325.
- [22] B. Liu, W. Hsu, Y. Ma, Mining association rules with multiple minimum supports, in: U. Fayyad, S. Chaudhuri, D. Madigan (Eds.), *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '99*, ACM Press, New York, New York, USA, 1999, pp. 337–341. doi:10.1145/312129.312274.
- [23] F. Z. El Mazouri, M. C. Abounaima, K. Zenkouar, Data mining combined to the multicriteria decision analysis for the improvement of road safety: case of france, *Journal of Big Data* 6 (2019). doi:10.1186/s40537-018-0165-0.
- [24] K. Bönisch, A. vom Brocke, S. Dippel, A. Grümpel, L. Hagemann, C. Jais, D. Lösel,

- A. Müller, S. Müller, A. Naya, H. Schrade, C. Späth, C. Velt, A. Wild, M. Lechner, Deutscher schweine-boniturschlüssel (dsbs), 2017. URL: [https://www.fli.de/fileadmin/FLI/ITT/Deutscher\\_Schweine\\_Boniturschluessel\\_2017-06-30\\_de.pdf](https://www.fli.de/fileadmin/FLI/ITT/Deutscher_Schweine_Boniturschluessel_2017-06-30_de.pdf).
- [25] R. Srikant, R. Agrawal, Mining generalized association rules, *Future Generation Computer Systems* 13 (1997) 161–180. doi:10.1016/S0167-739X(97)00019-8.
- [26] A. Famili, W. M. Shen, R. Weber, E. Simoudis, Data preprocessing and intelligent data analysis, *Intelligent data analysis* 1 (1997) 3–23.
- [27] R. Agrawal, T. Imielinski, A. Swami, Database mining: a performance perspective, *IEEE Transactions on Knowledge and Data Engineering* 5 (1993) 914–925. doi:10.1109/69.250074.
- [28] A. Vabalas, E. Gowen, E. Poliakoff, A. J. Casson, Machine learning algorithm validation with a limited sample size, *PloS one* 14 (2019) e0224365. doi:10.1371/journal.pone.0224365.
- [29] J. Han, J. Pei, Y. Yin, Mining frequent patterns without candidate generation, *ACM SIGMOD Record* 29 (2000) 1–12. doi:10.1145/335191.335372.
- [30] P. Fournier-Viger, T. Gueniche, S. Zida, V. S. Tseng, Erminer: sequential rule mining using equivalence classes, in: *International Symposium on Intelligent Data Analysis, 2014*, pp. 108–119.

## A. Discretization of continuous values

Data	Function	Unit	Ranges for a value $x$
Age (start of rearing)	–	kg	$x = 23; x = 24; \dots; 32 \leq x$
Birth weight	–	kg	$x < 0.5; 0.5 \leq x < 1.0; \dots; 4 \leq x$
Brightness	Min., Max., Mean, Diff.	lx	$x < 40; 40 \leq x < 80; \dots; 520 \leq x$
Control weight	–	kg	$x < 4; 4 \leq x < 6; \dots; 10 \leq x$
Daily increase	–	g	$x < 100; 100 \leq x < 125; \dots; 275 \leq x$
Food	Sum (daily)	kg	$x < 30; 30 \leq x < 40; \dots; 60 \leq x$
	Sum (weekly)	kg	$x < 180; 180 \leq x < 210; \dots; 360 \leq x$
	Diff. (weekly)	kg	$x < 30; 30 \leq x < 40; \dots; 60 \leq x$
Humidity	Min., Max., Mean, Diff.	%	$x < 30; 30 \leq x < 35; \dots; 90 \leq x$
Manipulable material	Sum (daily)	kg	$x < 0.75; 0.75 \leq x < 1.5; \dots; 6 \leq x$
	Sum (weekly)	kg	$x < 6; 6 \leq x < 7; \dots; 30 \leq x$
	Diff. (weekly)	kg	$x < 0.75; 0.75 \leq x < 1.5; \dots; 6 \leq x$
Temperatures (air, ground)	Min., Max., Mean	$^{\circ}C$	$x < 15; 15 \leq x < 18; \dots; 39 \leq x$
	Diff.	$^{\circ}C$	$x < 1; 1 \leq x < 2; \dots; 5 \leq x$
Temperatures (flow, return flow)	Min., Max., Mean	$^{\circ}C$	$x < 10; 10 \leq x < 15; \dots; 100 \leq x$
	Diff.	$^{\circ}C$	$x < -21; -21 \leq x < -18; \dots; 21 \leq x$
Water	Sum (hourly)	L	$x < 5; 5 \leq x < 10; \dots; 25 \leq x$
	Sum (daily)	L	$x < 10; 10 \leq x < 20; \dots; 250 \leq x$
	Sum (weekly)	L	$x < 50; 50 \leq x < 100; \dots; 500 \leq x$
	Diff. (weekly)	L	$x < 10; 10 \leq x < 20; \dots; 100 \leq x$
Weight (start of rearing)	–	kg	$x < 4; 4 \leq x < 6; \dots; 10 \leq x$