Data mining as a hatchery process evaluation tool

Daniela Regina Klein1*, Marcos Martinez do Vale3*, Mariana Fernandes Ribas da Silva3*, Micheli Faccin Kuhn1*, Tatiane Branco4*, Marcos Portella dos Santos5*, Mauricio Portella dos Santos1*

1Universidade Federal de Santa Maria – Depto. de Zootecnia, Av. Roraima, 1000 – 97105-000 – Santa Maria, RS – Brasil.
5Instituto de Zootecnia, Aparecida de Goiânia, Goiás, Brasil.

*Corresponding author <danniewk@yahoo.com.br>

Introduction

The expansion of poultry farming is founded on artificial incubation, one of the most important segments of the poultry chain, which ensures the performance of breeders and the quality of day-old chicks. With the evolution of available technology, the hatchery process now generates more data with potential for information extraction that can help increase process performance.

Classic data processing methods are at times not efficient in elaborating biological responses and factors (Mehri, 2013) such as genetic selection for meat or egg production, which influence the development and metabolism of the embryo during the hatching process (Buzała et al., 2015; Burggren et al., 2015).

A number of data and knowledge strategies to develop precision livestock farming are currently emerging such as the expert systems involving data mining (Vale et al., 2008), fuzzy logic (Sousa et al., 2016; Dominiaik and Kristensen, 2017) and artificial neural networking (ANN; Kumar and Hancle, 2015). The performance of ANN when applied to egg hatchability prediction is satisfactory (Bolzan et al., 2008), but it does not explain how the features are used to solve the problem. On the other hand, data mining (DM) explains the acquisition of knowledge in terms of logical rules, such as the construction of classification trees. DM has been successfully applied in animal science to animal environment (Vale et al., 2010), laying breeders’ production (Ferreira et al., 2013), breeding certification (Vieira et al., 2015), animal welfare (Moi et al., 2014), and egg quality prediction (Orhan et al., 2016).

Applying DM to organizing and analyzing a sizeable pool of data, the solutions are obtained faster than traditional methods with optimal predictive performance, which are needed in industrial applications (Koksal et al., 2011). Another advantage of the application of DM is that a mixture of numeric, categorical, and date data, can tolerate missing and noisy data, thereby increasing efficiency and detect potential quality problems in processes (Lei-da Chen and Frolick, 2000), including the poultry chain.

The aim of this study was to evaluate the applicability of the DM technique to building decision trees, data from hatchery results from broiler breeders and laying hen breeders housed under different systems.

Materials and Methods

The study used a database of hatcheries of brown egg laying hen breeders and broiler breeders from Apr to Sept 2011. Broiler breeders were housed in pens with shavings used for litters and reproductive management with natural mating. Laying hen breeders were housed in cages using artificial insemination. It was expected that the different systems would generate differences that the DM technique would probably explain.

The analysis used the DM technique for a classification task, having as classifier two genetic classes: broiler breeders (BB) and laying hen breeders (LH). The database had 49 egg hatching batches, 33 from LH and 16 from BB, and the attributes and distributions are in Table 1.
The database was analyzed using the Weka® software tool version 3.7.8, and the classification algorithm J48 for classification tasks. A classification tree, a graphical view of inverted tree form, was generated in which the root node is the first variable with the highest classification power presenting below the branches, formed by the other attributes that allow for classification.

The data mining technique was applied according to Vale et al. (2008) following the CRISP-DM (Cross Industry Standard Process for Data Mining) methodology. Three approaches to database analysis were used as follows: 1. analysis of database without attribute selection; 2. using attribute selection algorithms available in Weka®; and 3. attribute selection based on the help of experts with a minimum of two years of experience in hatchery.

The selection of attributes by Weka® algorithms seeks to find in databases those attributes which are determinants in the process of knowledge extraction. The algorithms used for the attribute selection were: InfoGain, which evaluates the value of an attribute by measuring the gain of information with respect to the class; CFS (Correlation-based Feature Selection), which evaluates the value of a subset of attributes by considering the individual predictive capacity of each resource along with an inherent degree of redundancy; and Gain-Ratio, which evaluates the value of an attribute by measuring the rate of gain with respect to the class (Witten et al., 2016).

The selection of the best model was based on the general precision of the models and the accuracy of the classes according to Vale et al. (2008) and Vale et al. (2010), using a contingency matrix and interpretation of the classification rules by experts.

The database was submitted to an analysis of variance, and the means between the classes were different ($p < 0.001$).

### Results

DM resulted in three classification trees of great relevance. The first classification modeled from the database without attribute selection [Figure 1, model precision of 76 %], classified the performance of hatching eggs from LH with better performance than those of BB, and class accuracy of 0.82 and 0.63 respectively.

The second classification approach used only relative values as directed by an expert. The percentage of fertile eggs was the main attribute able to classify the strains [Figure 2, model precision of 90 %, class accuracy of 0.94 and 0.82 respectively for LH and BB]. The model rules determined that the percentage of fertile eggs less or equal or are less than 95 % was the variable with the highest classification power to define LH. This result confirms the results of Figure 1.

The third tree using the Weka® attribute selection approach is shown in Figure 3 [94 % model precision, and class accuracy 1.00 and 0.84 respectively for LH and BB]. The percentage of fertile eggs was the only variable capable of classifying LH and BB egg performance.

The attributes of root nodes from the three decision trees were submitted to analysis of variance, and the means between the classes were different ($p < 0.001$).

### Table 1 – Database attribute distribution for laying hen breeders (LH) and broiler breeders (BB) hatchery process.

<table>
<thead>
<tr>
<th>Attributes</th>
<th>Mean</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Std. deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dbtemp incubator (°C)</td>
<td>36.24</td>
<td>35.07</td>
<td>37.09</td>
<td>0.58</td>
</tr>
<tr>
<td>Wbtemp incubator (°C)</td>
<td>31.62</td>
<td>30.44</td>
<td>32.43</td>
<td>0.75</td>
</tr>
<tr>
<td>Dbtemp hatcher (°C)</td>
<td>36.58</td>
<td>35.17</td>
<td>37.33</td>
<td>0.53</td>
</tr>
<tr>
<td>Wbtemp hatcher (°C)</td>
<td>31.14</td>
<td>29.17</td>
<td>34.25</td>
<td>1.20</td>
</tr>
<tr>
<td>Hatching eggs</td>
<td>1056</td>
<td>140</td>
<td>5678</td>
<td>923.57</td>
</tr>
<tr>
<td>Hatch rate (%)</td>
<td>75.40</td>
<td>44.47</td>
<td>92.61</td>
<td>10.89</td>
</tr>
<tr>
<td>Hatchability (%)</td>
<td>82.73</td>
<td>51.53</td>
<td>95.07</td>
<td>10.56</td>
</tr>
<tr>
<td>Fertile eggs (%)</td>
<td>91.21</td>
<td>74.19</td>
<td>99.21</td>
<td>6.62</td>
</tr>
<tr>
<td>Infertile eggs (%)</td>
<td>8.79</td>
<td>0.79</td>
<td>25.81</td>
<td>6.62</td>
</tr>
<tr>
<td>M1 (%)</td>
<td>2.34</td>
<td>0</td>
<td>5.78</td>
<td>1.29</td>
</tr>
<tr>
<td>M2 (%)</td>
<td>1.81</td>
<td>0</td>
<td>13.20</td>
<td>2.87</td>
</tr>
<tr>
<td>M3 (%)</td>
<td>0.92</td>
<td>0</td>
<td>10.84</td>
<td>1.68</td>
</tr>
<tr>
<td>M4 (%)</td>
<td>2.50</td>
<td>0</td>
<td>10.69</td>
<td>2.31</td>
</tr>
<tr>
<td>Pecked eggs (%)</td>
<td>2.32</td>
<td>0</td>
<td>13.31</td>
<td>3.00</td>
</tr>
<tr>
<td>Contaminated egg (%)</td>
<td>0.34</td>
<td>0</td>
<td>4.51</td>
<td>0.81</td>
</tr>
</tbody>
</table>

Wbtemp = Wet bulb temperature; Dbtemp = Dry bulb temperature; Embryonic mortality determined with embryo diagnosis (M1: in the first 48 h, M2: 3rd to 7th day, M3: 8th to 14th day, M4: 15th to 20th day).

### Figure 1 – Hatchery process classification from laying hen breeders (LH) and broiler breeders (BB) eggs. HRate = Hatching rate; HEgg = Hatched eggs; IRH = Incubator wet bulb temperature; Cont = Percentage of contaminated eggs; INF = Infertile eggs; FER = Fertile eggs.
constant, absolute egg number information may not be a good indicator of the incubation process, requiring relative values that can be preprocessed.

In the second tree BB has higher egg fertility and higher embryo mortality in the M2 phase, between the third and day of incubation (Buzala et al., 2015; Van Emous et al., 2015). This initial higher embryonic mortality is due to eggshell quality, egg size and, higher egg contamination from litter and nests (Ishaq et al., 2014; Parisi et al., 2015). The result reinforces the efficiency of the technique in extracting patterns for the differentiation of groups with different patterns.

The third classification tree focuses on the minimum information necessary to describe knowledge (Buczak and Guven, 2016). The rates of fertile eggs had great classification power and premise to identify classification rules linked with the production process (Vale et al., 2008).

The three approaches used in this DM confirm the importance of preprocessing data. Feature selection by experts or using algorithms is necessary in order to improve model precision and accuracy, removing irrelevant and redundant data (Kashef and Nezamabadi-Pour, 2015; Ramya et al., 2017).

The differences between the means of root node attributes of LH and BB prove that there are differences between the classes analyzed thereby improving the applicability of the DM technique and use of decision trees for extracting useful patterns from hatchery databases. Thus, taking note of other applications in animal science, the decision trees can be applied across the industry to quickly and efficiently find the patterns and solutions for quality problems.

**Conclusion**

The data mining technique allows for the classification of hatchery fertile eggs from different genetic groups. Hatching rates and the percentage of fertile eggs are the attributes with the highest classification power. Feature selection and the implementation of strategies based on the experience of experts improve model performance.

**Authors’ Contributions**


**Acknowledgments**

To the FAPERGS (Fundação de Amparo à Pesquisa do Estado do Rio Grande do Sul) for providing a Master's degree scholarship.
References


