Unsupervised outlier detection in questionnaire data

M.A. Krogh[1]*, P.S. Enemark[2], L. Foldager[1,3]

[1]Aarhus University, Department of Animal Science & [3]Bioinformatics Research Centre, Denmark, [2]Arla Foods, Denmark

*Corresponding author: mogenskrogh@anis.au.dk

Unsupervised outlier detection described here can be used to:

- **Rank questions with most outlierness information**
- **Rank herds according to outlierness**
- **Highlight the possible problematic answers**

Introduction

Arlagaarden®Plus is a voluntary information collection system, where dairy farmers can choose to make information about their production system and management practices available for Arla. One part of Arlagaarden®Plus is a large questionnaire consisting of 90+ different questions that the farmers need to answer or update quarterly.

The purpose of this study is to develop a method for detection of questionnaires that contain answers that diverge from the majority (outliers) that can focus subsequent quality control of the answers.

Materials & Methods

- 8,501 completed questionnaires from Dec. 2018
- Herds in Belgium, Denmark, Germany, Great Britain, Luxembourg, the Netherlands and Sweden
- Electronic questionnaire presented in native langue
- Questionnaire answers mostly categorical

An outlier detection algorithm described by Pang et al. (2016)[1] was chosen. The method which is called Coupled Biased Random Walk (CBRW), estimates the outlierness – the chance of being an outlier – of each answer by capturing both intra- and inter-question answer couplings.

Results

The results from implementing the algorithm on 15 questions related to calves can be seen in Table 1. The results show that 'Age where calves are given milk replacer', 'Weaning age', 'Herd size of cows' and 'Herd size of calves' are the most important questions in detection of outliers (yellow) whereas 'Calving pen' and 'Heifer hotel' are almost independent of the answers to other questions (very light yellow).

The three answers within each herd, that have the largest chance of being outliers or include possible wrong answers are highlighted in light purple. For the three herds highest outlier score rank, it seems that the combination of answers to 'Age where calves are given milk replacer' and 'Weaning age' is unlikely if the herd size is large (more than 400 cows).

The results demonstrate that the CBRW-algorithm can be used for this purpose, but requires that wrong answers are rare events and the questions are correlated.

[1]Pang, Guansong & Chen "Unsupervised feature selection for outlier detection by modelling hierarchical value-feature couplings." 2016 IEEE 16th International Conference on Data Mining (ICDM). IEEE, 2016.

Table 1: Ranking of 7 herds in terms of outlierness (outlier score rank) in a part of the questionnaire about calves. The three answers that are most likely to be outliers for each herd are marked in light purple. The yellow color gradient shows the question's ability to detect outliers, where light yellow indicate poor ability to detect outliers.

Outlier score rank	Calving pen	Heifer hotel	Bull calves	Calf hut	Calf single pen	Calf group age (wk)	Calf stable with cow stable	Feeding system	Milktype	Milk replacer age (wk)	Weaning age (wk)	Salmonella	BVD	Herd size cows	Herd size calves
1	Yes	Yes	Euthanized	Yes	No	3	Yes	Automatic	Milk replacer	3	11	No	Yes	Large	Large
2	Yes	Yes	Euthanized	Yes	Yes	6	Yes	Bucket	Raw milk & replacer	4	11	Yes	Yes	Large	Large
3	Yes	No	Euthanized	Yes	Yes	5	Yes	Bucket	Pasteurized raw milk	3	9	Yes	No	Large	Large
…	…	….	…	…	…	…	…	…	….	…	…	…	…	…	….
501	Yes	Yes	Sold to other farm	No	No	1	Yes	Automatic	Milk replacer	1	9	No	No	Large	Small
502	Yes	Yes	Sold to other farm	No	Yes	8	No	Bucket	Raw milk & replacer	4	20	Yes	Yes	Small	Small
…	…	…	….	…	…	…	…	….	…	…	…	….	…	…	…
5001	Yes	Yes	Sold to other farm	Yes	Yes	2	Yes	Bucket	Raw milk & replacer	1	8	Yes	Yes	Medium	Medium
5002	Yes	Yes	Live stock market	No	Yes	6	No	Bucket	Raw milk	6	8	No	No	Medium	Small