Imprecise Classification of the Gram Status of the Causal Pathogen of Clinical Mastitis

Meizhu Li, Jasper De Bock & Gert de Cooman

IDlab, Ghent University, Belgium

{meizhu.li, jasper.debock, gert.decooman}@UGent.be

GHENT UNIVERSITY

Introduction

Clinical Mastitis (CM) – a type of udder infection – is one of the most frequent and costly diseases in a dairy herd. Its main characteristics are:

- reduction in a cow's milk production and overall yield;
- painful for the infected cows;
- can be transmitted between cows.

→ An *effective* and *quick* CM treatment is of economic importance.

Treatment Appropriate treatment is selected based on the gram status of the causal pathogen.

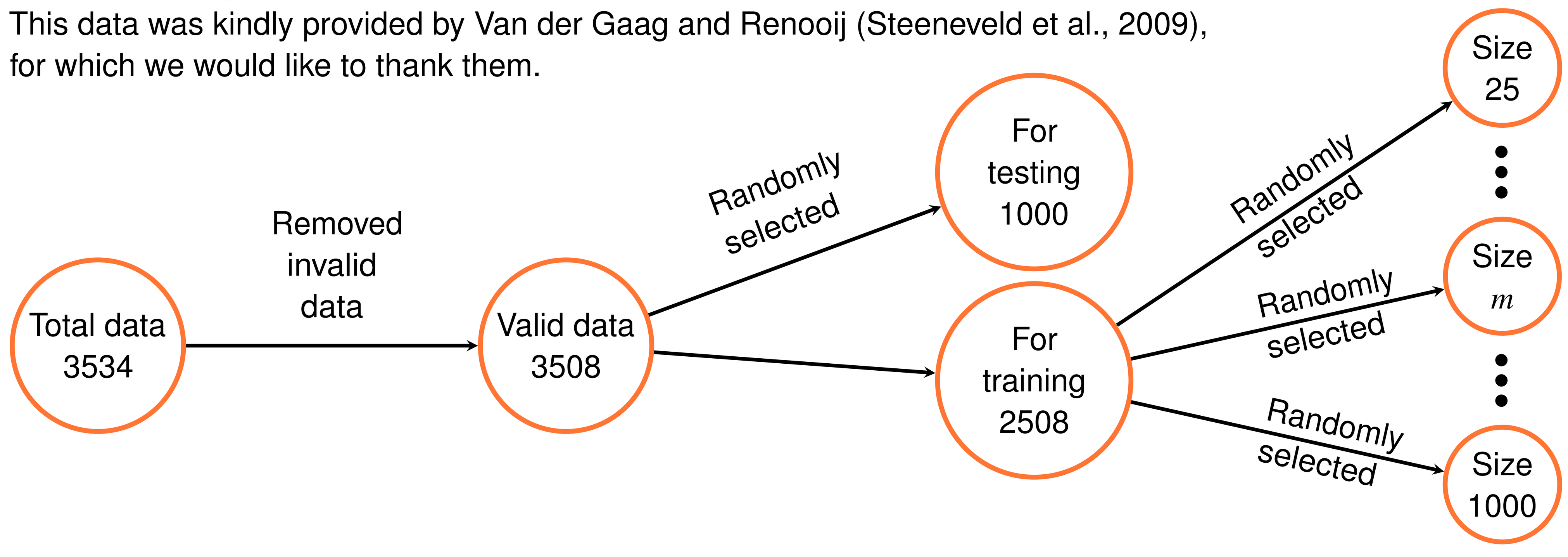
- *Bacteriological culturing* is ineffective due to long waiting times.
- *Classifiers* help farmers to determine the gram status of the causal pathogen, allowing them to select an appropriate treatment.

Motivation for imprecision

- If a single gram status of the causal pathogen is convincingly established, a narrow-spectrum antibiotic is better.
- If the gram status of the causal pathogen cannot be convincingly established, a broad-spectrum antibiotic is more appropriate.

Data Setting

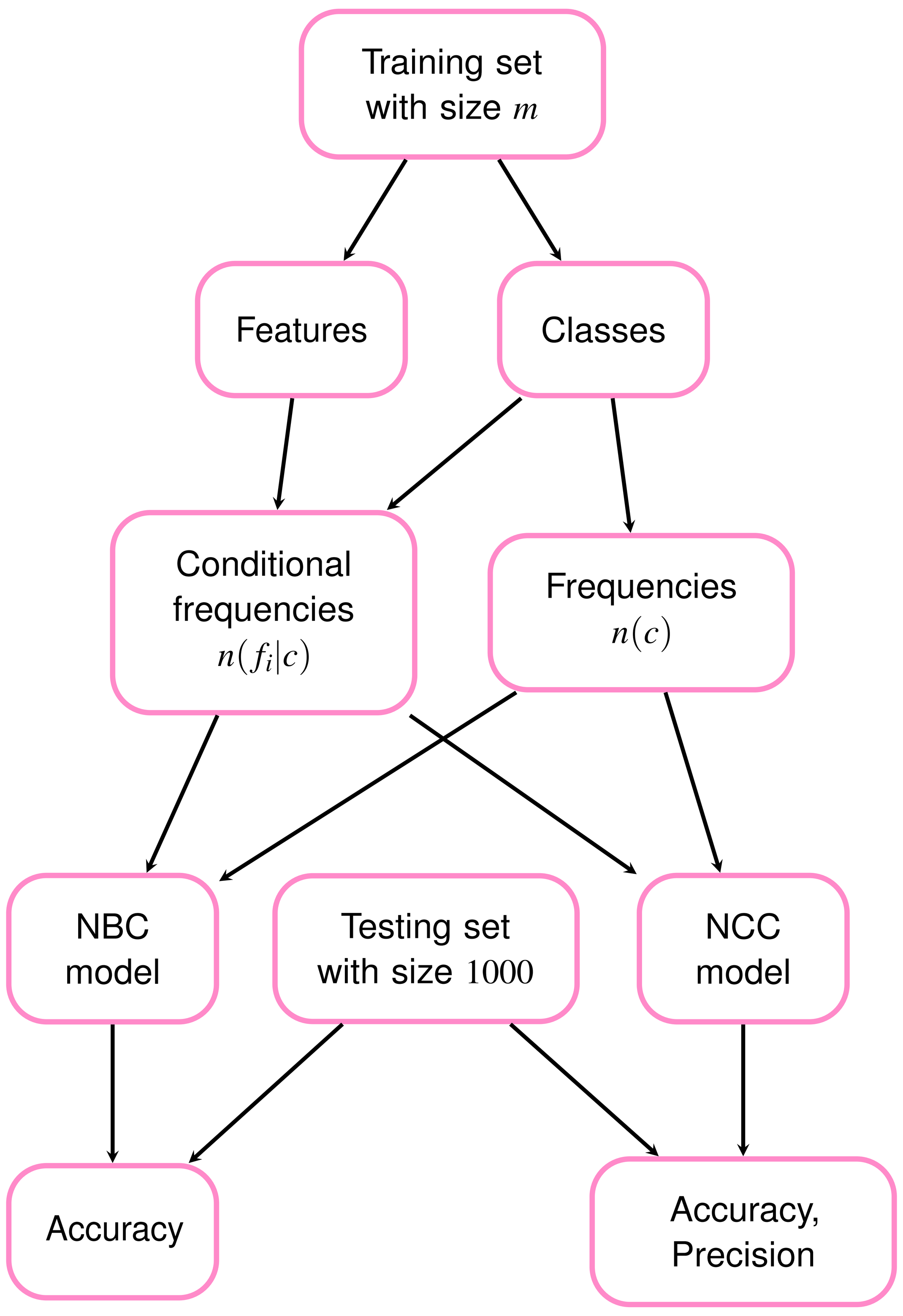
For constructing classifiers, a set of 3534 clinical mastitis cases was used. This data was kindly provided by Van der Gaag and Renooij (Steeneveld et al., 2009), for which we would like to thank them.



Four features were selected from the dataset in this case study – TEXT, GRAM1, GRAM2 and SICK. The class that needs to be determined is GRAM.

Name	Classes	Classifications	Description
GRAM	2	Gram-negative, Gram-positive	Gram status of current CM case
TEXT	5	Normal, small flakes, big flakes, serous, viscous	Texture of the milk of cow with CM
GRAM1	3	No previous DM, gram-positive, gram-negative	Gram history 1-30 d before current CM
GRAM2	3	No previous DM, gram-positive, gram-negative	Gram history >30 d before current CM
SICK	2	Not sick, sick	Cow sick at moment of CM

Flow Chart



Models

Imprecise model We use the Naive Credal Classifier (NCC) (Marco Zaffalon, 2001) as the imprecise model. Let g_+ and g_- denote the classes of gram-positive and gram-nagative. $(f_1,\dots,f_k)=\mathbf{f}$ is a vector of k feature variables in this case. For any $c\in\{g_+,g_-\}$, the conditional lower and upper probabililties of c given $\mathbf{f}$ are:

$$\overline{p}(c|\mathbf{f}) := \frac{\prod_{i=1}^{k}[\overline{p}(f_i|c)]\overline{p}(c)}{\prod_{i=1}^{k}[\overline{p}(f_i|c)]\overline{p}(c)+\prod_{i=1}^{k}[\underline{p}(f_i|c^c)]\underline{p}(c^c)},$$

$$\underline{p}(c|\mathbf{f}) := \frac{\prod_{i=1}^{k}[\underline{p}(f_i|c)]\underline{p}(c)}{\prod_{i=1}^{k}[\underline{p}(f_i|c)]\underline{p}(c)+\prod_{i=1}^{k}[\overline{p}(f_i|c^c)]\overline{p}(c^c)},$$

where $c^c := \{g_+,g_-\}\setminus c$, and where

$$\underline{p}(c) := \frac{n(c)}{N+s} \quad and \quad \overline{p}(c) := \frac{n(c)+s}{N+s},$$

$$\underline{p}(f_i|c) := \frac{n(f_i|c)}{n(c)+s} \quad and \quad \overline{p}(f_i|c) := \frac{n(f_i|c)+s}{n(c)+s}$$

where N is the total number of training units, $n(c)$ and $n(f_i|c)$ are the observed frequencies of class c and of $(f_i|c)$ in the N observations. We choose s equal to 1.

Precise model We use the Naive Bayes Classifier (NBC), which is well known in classification, as the precise model.

Model measurement TP and FP denote the number of true and false positive cases, TN and FN denote the number of true and false negative cases. UP and UN is the number of positive and negative cases for which the model gives a result of 'I don't know'.

Data \ Classifier	g_+	g_-	*I don't know*
g_+	TP	FN	UP
g_-	FP	TN	UN

Accuracy and Precision are used to analyse the performance of the NBC and NCC classifiers. The values of precision for the NBC classifier are always equal to 1.

$$\text{Accuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}}$$

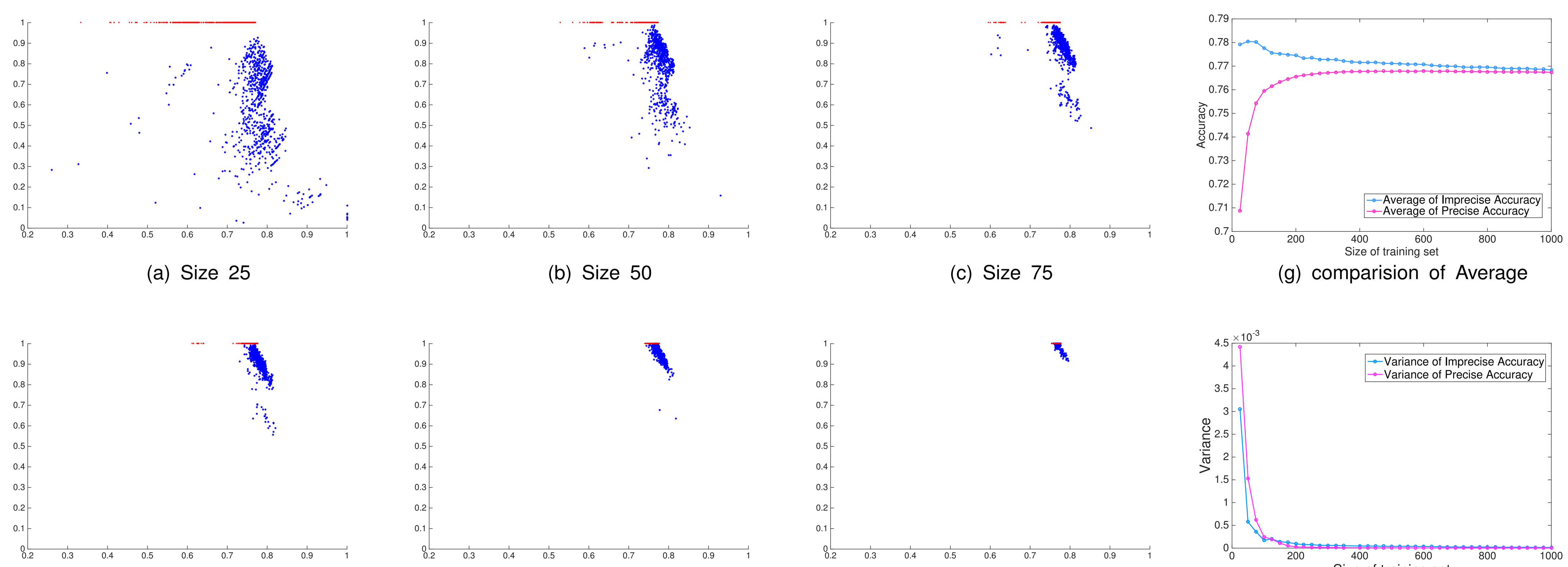
$$\text{Precision} = \frac{\text{TP}+\text{TN}+\text{FP}+\text{FN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}+\text{UP}+\text{UN}}$$

Decision rules In the NBC, we select the class with largest probability as the output. In the NCC, the decision rule is shown in following table:

Condition	Output
$\overline{p}(g_-)\le\underline{p}(g_+)$	Class g_+
$\overline{p}(g_+)\le\underline{p}(g_-)$	Class g_-
$\overline{p}(g_+)>\underline{p}(g_-)$ and $\overline{p}(g_-)>\underline{p}(g_+)$	'I don't know'

Results for Different Sizes of Training Sets

For every size of training set, 1000 precise and imprecise classifiers were constructed for random datasets of that size. For the NCC (blue) and NBC (red), the results are depicted in graphs (a) – (f). The x-axis depicts the Accuracy and the y-axis depicts the precision. Graph (g) and (h) compare the average and variance of the accuracy between both models.



(a) Size 25 (b) Size 50 (c) Size 75 (g) comparision of Average

(d) Size 100 (e) Size 200 (f) Size 500 (h) comparision of Variance

Conclusion and Future Work

Conclusion

- In the NCC, the precision depends strongly on the training set, whereas the accuracy seems more stable across different training sets. In the NBC, the precision is always 1, but the accuracy varies greatly across different training sets.
- The imprecise model has a better accuracy than the precise one, at the cost of a loss of precision.
- For small training sets, the imprecise model has a lower variance than the precise one.

Future work

- Investigate and compare several imprecise classification methods by applying them to our case study on clinical mastitis.
- Compare the obtained results with other precise methods.
- Study the effect of the parameter s on the performance of the NCC model.