

## Rule discovery for region linking: Preliminary results

**Objective:** Find true links between region X in one view and region Y in the other view for a particular breast, based on:

1. Correspondence score of the link X-Y (*CorrScore*).
2. Absolute difference between the false positive levels of regions X and Y (*Fplevel\_diff*).
3. Individual false-positive (FP) levels computed from the CAD system (*FPLLevel*).

### Data description:

*Raw data representation obtained from the CAD system:* Example of one row

| <i>ImageID</i>                  | <i>Current view region X</i> | <i>X- True positive?</i> | <i>Other view region Y</i> | <i>Y- True positive?</i> | <i>Y - xLocation</i> | <i>Y - yLocation</i> | <i>Corr. score X-Y</i> | <i>Corr. score &gt;0.5?</i> | <i>Fplevel of Y</i> | <i>Fplevel_diff X-Y</i> |
|---------------------------------|------------------------------|--------------------------|----------------------------|--------------------------|----------------------|----------------------|------------------------|-----------------------------|---------------------|-------------------------|
| 2093376mr<br>(m=MLO<br>r=right) | 0                            | 1                        | 4                          | 0                        | 579                  | 354                  | 0.2868                 | 0                           | 1.4                 | 2.88                    |

### Data re-formatting:

Instead of considering both views for a particular breast separately and try to link the regions between them, we look at the breast as whole. Thus, for a particular breast we have regions whose number equals the total number of regions from both MLO and CC view. For example, if we have 3 regions in MLO view and 5 regions in CC view, we have in total for the breast 8 regions. Between these regions there are two-way links having different likelihoods for true linking, which are computed from the CAD system.

- *Class variable: Exam\_Breast*, i.e., for a particular exam we consider each of the breast separately.
- *Number of classes: 2*
  - positive breasts: 339
  - negative breasts: 1555
- *Number of instances:*
  - regions: 18 725
  - links: 92 582

**Methodology:** Probabilistic rule discovery based on nFOIL (Inductive Logic Programming and Naïve Bayes classifier). The idea is to find rules for correctly linking true regions. The rules with the highest probability for correct classification are selected.

### Define predicates:

```
class_values(positive, negative);  
class(Exam_Breast);  
region(Exam_Breast, RegionID, True, FPLLevel)  
linkLik(Exam_Breast, RegionID1, RegionID2, TrueLink, CorrScore, Fplevel_diff)
```

*True* is 1 if a region is positive; otherwise it is 0.

*TrueLink* is 1 if both linked regions are positive (i.e., 1-1 link); otherwise it is 0.

Goal: Based on the defined predicates, find rules that cover as much as possible positive examples (breasts) and do not cover negative ones.

**Results**: Model learned with 2-fold cross-validation (CV Error = 0.0174 (33/1894)):

region(A,B,1,0.01\_0.1) linkLik(A,C,B,1,0.8\_1.0,0.01\_0.1) (116)  
region(A,B,1,0.1\_0.2) linkLik(A,C,B,1,.0.8\_1.0,0.1\_0.2) (36)  
region(A,B,1,0.2\_0.4) (39)  
region(A,B,1,0.01\_0.1) linkLik(A,C,B,1,0.6\_0.8,0.01\_0.1) (15)  
region(A,B,1,0.6\_0.8) (17)  
region(A,B,1,1\_1.2) (10)  
region(A,B,1,1.4\_1.6) (9)  
region(A,B,1,0\_0.01) (8)  
region(A,B,1,3.6\_3.8) (5)  
region(A,B,1,4.2\_4.4) (4)

Note that each row represents a rule and the number in the brackets is the number of positive examples covered by the rule per fold.

### Interpretation of the results:

**region(A,B,1,0.01\_0.1)** means that for exam\_breast A, we have B that is a positive region with FPlevel between 0.01 and 0.1. For example A=2035104right and B = region1, i.e., A is a positive breast and B is the true region

**linkLik(A,C,B,1,0.8\_1.0,0.01\_0.1)** means that for exam\_breast A, regions C and B are truly linked (1) with likelihood between 0.8 and 1 and the absolute difference between the FP levels of the regions C and B is between 0.01 and 0.1.

Hence, for example the first rule must be read as follows:

**region(A,B,1,0.01\_0.1) linkLik(A,C,B,1,0.8\_1.0,0.01\_0.1)**

If region the true region B with FPlevel between 0.01 and 0.1 is linked to another true region from the same exam and breast with likelihood between 0.8 and 1 and the absolute FPlevel difference is between 0.01 and 0.1, then this is a true link and the breast is positive with a certain probability.

### Future work:

1. Add the absolute difference between the correspondence scores of the links X-Y and Y-X and use it to refine the rules derived.
2. Use x- and y-location of the region as additional information for refinement.
3. Use the derived rules to compute a new value for the probability that a link is true.