

In this document we provide a preliminary account on the definition of the (practical) problem we will concentrate now and the initial ideas on how we intend to tackle it.

## 1 What

Here we describe the database we will explore (in terms of first-order logic syntax) and a preliminary description of the problem. Data represented by predicates:

$image(ImageID, ExamID, Breast, View)$

$region(RegionID, ImageID, RegionN, Xloc, Yloc, GTRTruth, FPlevel)$

$hasRegion(I, X) : \neg region(R, I, X, XL, YL, C, FP)$

$linkLik(RegionID1, RegionID2, Likelihood, TrueLink)$

$linked(E, X, Y) : \neg image(I1, E, B, m), hasRegion(I1, X),$   
 $image(I2, E, B, c), hasRegion(I2, Y), linkLik(X, Y, L, 1)$

Notes:

- ImageID is a consecutive number, which will uniquely identify the combination of exam, breast and view
- ExamID is the number in the first column in the current data set
- RegionID is a consecutive number in the table of regions, whereas RegionN is the number of the region ranging from 0 to 4 (see the data set) for the particular ImageID.
- GTRTruth takes 0 if region is not true cancer and 1 otherwise
- FPlevel is the false positive level computed by the CAD system (needs to be added to the current data)
- linkLik is the likelihood for a link between two regions given in the last but one column in the current data
- TrueLink takes 0 if it is a false link and 1 otherwise

Problem:

For given exam  $E$ , find the probability that region  $X$  from one view and region  $Y$  from the other view represent one finding  $P(linked(E, X, Y))$  given:

1. the likelihood of the linking of  $X$  to other two regions in the other view
2. the likelihood of linking of  $Y$  to  $X$ .
3. the individual FP levels (and peak levels) computed from the CAD system.

## 2 How

One possible method for knowledge acquisition through automatic rule induction from database:

**Inductive logic programming + NB/TAN:** On the one hand, in relational learning or inductive logic programming, a set of rules (or clauses) is induced. This set is a collection of disjunctive hypotheses: an instance is classified as positive if it satisfies the conditions present in one of the rules. On the other hand, we have that a probabilistic model defines a joint probability distribution over a class variable and a set of attributes/features. By integrating this two approaches we can have clauses as “attributes” over which a joint probability distribution is defined. When using Naïve Bayes (as the probabilistic model), this means that “clauses are independent”. In short, the idea is that an ILP algorithm learns a rule, which is combined through the use of a classifier, such as a naïve Bayes. Here the rules are scored by how much they improve the classifier: if a rule improves the classifier, then it is retained; otherwise, it is discarded. As stated by Burnside *et al.*, the advantages of ILP (over other data mining techniques) are the use of background knowledge to narrow the search space and return of humanly comprehensible results.

### 2.1 ILP, SRL and Bayes Classifiers

Inductive Logic Programming (ILP) provides algorithms for learning hypotheses, which are expressed as logical rules, from data. Given a language specification  $L$  (for instance, Horn clauses), which specifies how hypotheses are constructed; a knowledge base  $B$ , such as a database; and, a (finite) set of examples, the aim of an ILP system is to find a set  $H$  of hypotheses (or rules) that cover all positive examples and no negative examples. An optional set of constraints on acceptable hypotheses can also be given. “The key question is how rules can be combined in order to obtain a useful classifier.” [1].

As stated by Burnside *et al.*, the advantages of ILP (over other data mining techniques) are the use of background knowledge to narrow the search space and return of humanly comprehensible results. Discovering new knowledge, through non-obvious relations among objects, are of particular interest - specially, when referring to structured data.

Statistical Relational Learning (SRL), as the name indicates, aims at learning statistical models from (relational) data. “SRL advances beyond Bayesian network learning and related techniques by handling domains with multiple tables, representing relationships between different rows of the same table, and integrating data from several distinct databases.” [2].

*View learning*, in particular, shows how SRL algorithms can benefit from the ability to define new views, i.e., new fields defined in terms of existing fields and background knowledge. In this case, a view can represent connections between related examples. In this context, ILP algorithms can be used in order to learn rules that define this new database view or fields.

The approach suggested by Burnside *et al.* (SAYU), the same way as the system developed by Kersting *et al.* (nFOIL) [3], uses an intertwined execution of an ILP algorithm and a Bayes classifier. An ILP algorithm learns a rule, which is combined through the use of a classifier, such as a naïve Bayes. Here the rules are scored by how much they improve the classifier: if a rule improves the classification, it is retained; otherwise, it is discarded, resulting in the old classifier being restated.

## 2.2 Development phase

- Use previously developed systems, such as nFOIL [3].
- Develop own system by using Bayesian networks Toolbox in Matlab (and/or Prolog).
- If possible, compare the systems.

## References

- [1] J. Davis, E. Burnside, I. Dutra, D. Page, and V. S. Costa. An integrated approach to learning bayesian networks of rules. In *Proceedings of the European Conference of Machine Learning (ECML)*, 2005.
- [2] J. Davis, E. Burnside, I. Dutra, D. Page, R. Ramakrishnan, V. S. Costa, and J. Shavlik. View learning for statistical relational learning: With an application to mammography. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence (IJCAI)*, 2005.
- [3] N. Landwehr, K. Kersting, and L. de Raedt. nfoil: Integrating naïve bayes and FOIL. In M. Veloso and S. Kambhampati, editors, *Proceedings of the Twentieth National Conference on Artificial Intelligence (AAAI-05)*, pages 795–800, USA, 2005.