On Applying Probabilistic Logic Programming to Breast Cancer Data

Joana Côrte-Real✉, Inês Dutra, and Ricardo Rocha

CRACS & INESC TEC and Faculty of Sciences, University of Porto
Rua do Campo Alegre, 1021/1055, 4169-007 Porto, Portugal
{jcr,ines,ricroc}@dcc.fc.up.pt

Abstract. Medical data is particularly interesting as a subject for relational data mining due to the complex interactions which exist between different entities. Furthermore, the ambiguity of medical imaging causes interpretation to be complex and error-prone, and thus particularly amenable to improvement through automated decision support. Probabilistic Inductive Logic Programming (PILP) is a particularly well-suited tool for this task, since it makes it possible to combine the relational nature of this field with the ambiguity inherent in human interpretation of medical imaging. This work presents a PILP setting for breast cancer data, where several clinical and demographic variables were collected retrospectively, and new probabilistic variables and rules reflecting domain knowledge were introduced. A PILP predictive model was built automatically from this data and experiments show that it can not only match the predictions of a team of experts in the area, but also consistently reduce the error rate of malignancy prediction, when compared to other non-relational techniques.

1 Introduction

Probabilistic Inductive Logic Programming (PILP) is a subset of Statistical Relational Learning (SRL) that handles statistical information by using a probabilistic first-order logic language to represent data and their induced models. This technique merges technologies from the SRL and Inductive Logic Programming (ILP) [19] fields in order to automatically compose theories as understandable First Order Logic (FOL) sentences based on data annotated with probabilistic information. PILP manipulates structured representations of data so as to capture the logic relations that lie beyond the low-level features and reason about them by learning the (logical) structure of the data inductively.

The unique ability to combine the expressiveness of FOL rules with a degree of uncertainty makes PILP methods particularly well-suited to be applied in medical domains. Expert knowledge regarding the problem setting can be coded as facts or rules with varying frequencies or degrees of belief [15], and subsequently be used during the knowledge extraction stage to generate the final model. In addition, this final model also consists of a FOL theory which explains the behaviour of the system, and is easily interpretable by human experts (even though it may also be used to perform prediction over new examples).

Breast cancer is one of the most common forms of cancer and mammograms are the most commonly used technique to detect patients at risk. Image-guided core needle biopsy of the breast is then performed to decide on surgery. Biopsy is a necessary, but also aggressive, high-stakes procedure. The assessment of malignancy risk following breast core biopsy is imperfect and biopsies can be *non-definitive* in 5-15% of cases [2]. In particular, the dataset used in this work consists of demographic-related variables and information about the biopsy procedure and BI-RADS (Breast Imaging Reporting and Data System) [12] annotations, as well as domain knowledge annotated both prospectively and retrospectively by experts of three different areas: mammography, biopsy surgery and biopsy pathology. Using an automated decision support system is conducive to rigorous and accurate risk estimation of rare events and has the potential to enhance clinician decision-making and provide the opportunity for shared decision making with patients in order to personalize and strategically target health care interventions.

This work proposes a PILP decision support system targeted to this breast cancer setting. Contrary to other decision support systems, well-known in the literature (for example, Bayesian-based or SVM-based), the model proposed in this work combines probabilistic data with first order logic in order to produce both probabilistic outputs and human interpretable rules. The proposed setting includes experts' domain knowledge as (i) probabilistic rules in the background and (ii) probabilistic target values for examples. Experiments show that incorporating this domain knowledge in the model results in automated predictions which are statistically similar to those of a multidisciplinary team of human experts. Furthermore, the rules produced by the decision support system are human interpretable and relevant to the domain, which can help clinicians assess new cases.

2 Probabilistic Inductive Logic Programming

Introducing probabilistic information in a FOL setting allows for modelling facts or rules which are believed to be true to some degree or with a given frequency (as opposed to crisp true or false statements), which results in a closer representation of reality. Probabilities in a logic setting can also be used in cases where all the data were not gathered, since rules containing some information (if available from other sources) can be taken into account when building the final theory model. Additionally, in cases where there are privacy concerns, a similar approach can be used to avoid using the patient instances explicitly, while still considering some of the information contained in the original data.

More formally, PILP is a machine learning technique which learns predictive models from a set of probabilistic logic facts and rules. Like ILP, PILP uses a set of Probabilistic Examples (PE) and additional probabilistic logical information about the domain, the Probabilistic Background Knowledge (PBK), to find a model that explains the probabilistic examples. The PBK is a description of observed data composed of Horn clauses that can be annotated with probabilistic information known a priori. If not annotated, it is assumed that their probabilistic value is 1. The PE represent the observations that the system is attempting to explain. They also have probabilistic values a priori. Good models will approximate the probabilistic examples values with minimum error.

In this work, probabilities are annotated according to ProbLog's syntax, using *possible world semantics* [11]. Each fact $p_j :: c_j$ in the PBK represents an independent binary random variable in ProbLog, meaning that it can either be true with probability p_j or false with probability $1 - p_j$. This means that each probabilistic fact introduces a probabilistic choice in the model. Each set of possible choices over all facts of the PBK represents a possible world ω_i, where ω_i^+ is the set of facts that are true in that particular world, and $\omega_i^- = \omega_i \setminus \omega_i^+$ is the set of facts that are false. Since these facts have a probabilistic value, a ProbLog program defining a probabilistic distribution over the possible worlds can be formalized as shown in Equation 1.

$$P(\omega_i) = \prod_{c_j \in \omega_i^+} p_j \prod_{c_j \in \omega_i^-} (1 - p_j) \tag{1}$$

A ProbLog *query* q is said to be true in all worlds w^q where $w^q \models q$, and false in all other worlds. As such, the *success probability* of a query is given by the sum of the probabilities of all worlds where it is found to be true, as denoted in Equation 2.

$$P(q) = \sum_{\omega_i \models q} P(w_i) \tag{2}$$

PILP systems learn models in the form of probabilistic logic programs.

The theories used to explain examples in PILP are built from the literals that are present in the program's PBK. The rule (AND) search space is composed by all *Rules* whose body contains one or more of those literals. Rules can be combined using logical conjunction to form longer more *specific* rules. Let *Literals* be the set of distinct literals in the PBK. The AND search space is then the power set of *Literals*, $\mathcal{P}(Literals)$.

The theory (OR) search space can be defined in a similar way. Theories are formed by combining a set of distinct rules using logical disjunction. In the same way that literals are the building blocks of rules, rules are the building blocks of theories. Adding a rule to a theory makes it more general. The OR search space is then the set of all theories *Theories* such that $Theories = \mathcal{P}(Rules)$.

Fully exploring the PILP search space is equivalent to evaluating all theories in order to determine the best theory according to a given metric. This can be done in two steps: (i) exploring the AND search space, and (ii) exploring the OR search space. Algorithm 1 presents this procedure.

Algorithm 1 explores the AND search space in a direction of increasing specificity. It starts out by generating rules containing only one literal, using the mode declarations (line 2), and then uses these rules to generate combinations, which are possible according to the language bias, for the next iteration (lines 5–8), and removing the redundant rules. The combination process is repeated

Algorithm 1 $PILP\_search\_space(PBK, PE, MaxRuleLen, MaxTheoryLen)$

```
 1: $R_{all} = \emptyset$
 2: $R_1 = generate\_rules\_one\_literal(PBK, PE)$
 3: $R_{new} = R_1$
 4: $R_{len} = 1$
 5: while $R_{new} \neq \emptyset$ and $R_{Len} \leq MaxRuleLen$ do
 6:     $R_{all} = R_{all} \cup R_{new}$
 7:     $R_{new} = \{r_1 \wedge r_{new} \mid (r_1, r_{new}) \in R_1 \times R_{new}\}$
 8:     $R_{len} = R_{len} + 1$
 9: $T_{all} = \emptyset$
10: $T_1 = R_{all}$
11: $T_{new} = T_1$
12: $T_{len} = 1$
13: while $T_{new} \neq \emptyset$ and $T_{Len} \leq MaxTheoryLen$ do
14:     $T_{all} = T_{all} \cup T_{new}$
15:     $T_{new} = \{t_1 \vee t_{new} \mid (t_1, t_{new}) \in T_1 \times T_{new}\}$
16:     $T_{len} = T_{len} + 1$
17: return $T_{all}$
```

until it yields no new rules, or until the number of literals in the rules is greater than a pre-defined maximum number of literals. The set of initial theories T_1 is then populated with all rules in R_{all} (line 10). Similarly to the AND search space, T_1 is used to generate new theories T_{new} through combination using logical disjunction (lines 13–16). This process is analogous to the exploration of the AND search space.

3 Methodology

Breast cancer is one of the most common forms of cancer. Mammograms are the most commonly used technique to detect patients at risk. Image-guided core needle biopsy of the breast is then performed to decide on surgery. Biopsy is a necessary, but also aggressive, high-stakes procedure. The assessment of malignancy risk following breast core biopsy is imperfect and biopsies can be *non-definitive* in 5-15% of cases [2,3,4,14,17,18].

A non-definitive result means that the chance of malignancy remains high due to possible sampling error (i.e., the obtained biopsy is not representative of the suspicious finding), for which surgical excisional biopsy or aggressive radiologic follow-up is proposed. Non-definitive biopsies may therefore result in missed breast cancers (false negatives) and unnecessary interventions (false positives). In the US, the women over the age of 20 years have an annual breast biopsy utilization rate of 62.6 per 10,000 women, translating to over 700,000 women undergoing breast core biopsy in 2012. As a result of non-definitive biopsies, approximately 35,000-105,000 of these women will require additional biopsies or follow-up secondary to judged inadequacy of breast core biopsy.

Interpretation can be complex and error-prone, and thus particularly amenable to improvement through automated decision support, where rigorous and accurate risk estimation of rare events have the potential to enhance clinician decision-making and provide the opportunity for shared decision making with patients in order to personalize and strategically target health care interventions.

The dataset used for this experiment contains anonymised data from 130 biopsies dating from January 2006 to December 2011, collected from the School of Medicine and Public Health of the University of Wisconsin-Madison. The data was prospectively given a non-definitive diagnosis at radiologic-histologic correlation conferences. 21 cases were determined to be malignant after surgery, and the remaining 109 proved to be benign. For all of these cases, several sources of variables were systematically collected including variables related to demographic and historical patient information (age, personal history, family history, etc.), mammographic BI-RADS descriptors (like mass shape, mass margins or calcifications), pathological information after biopsy (type of disease, if it is incidental or not, number of foci, and so on), biopsy procedure information (such as needle gauge, type of procedure), and other relevant facts about the patient.

Probabilistic data was then added to (i) the Probabilistic Examples (PE) and (ii) the Probabilistic Background Knowledge (PBK). In the first instance, the confidence in malignancy for each case (before excision) is associated with the target predicate `is_malignant/1`. The chance of malignancy is an empirical confidence value assigned by a multidisciplinary group of physicians who meet to discuss and reach an agreement about each case. Thus, the target probabilities of examples represent the perceived chance of malignancy for each patient. A high probability indicates the team of physicians thinks the case is most likely malignant, and conversely a low probability indicates the case is most likely benign. This probabilistic value was then added to the probabilistic examples and a sample of the PE is presented next:

```
example(is_malignant(case1), 0.10).
example(is_malignant(case2), 0.15).
example(is_malignant(case3), 0.01).
```

Each example is a patient case and the three examples above are part of the PE used in this experiment (one per line). Each example has two arguments, the first being the target predicate `is_malignant/1` concerning a particular case (`case1`, `case2`, or `case3`) and the second the chance of malignancy of this case (10% for `case1`, 15% for `case2`, and 1% for `case3`).

Regarding the domain knowledge incorporated in the PBK, breast cancer literature values were used to complement the information on the characteristics of masses, since physicians rely on these values to perform a diagnosis. For example, it is well known among radiology experts in mammography that if a mass has a spiculated margin, the probability that the associated finding is malignant is around 90%. The same kind of information is available in the literature for mass shape or mass density (all part of the BIRADS terms). Figures 1, 2, and 3 show how these variables are encoded in the PBK, (the notation is `probability_value::relation(...)...`). Figure 1 encodes the probabilistic information regarding mass shape obtained from the literature. There are three possible rules, each one applicable to a particular kind of shape (oval, round, or irregular). A rule of this type can be read as *IF this Case has a Mass AND the Mass is of type Shape THEN this feature exists with probability P*. The probability value annotated in each rule is the frequency with which a mass whose shape is of that type is malignant. Independent rules such as the ones presented in Fig. 1 are not mutually exclusive. This means that a finding may have simultaneously an oval and round mass shape, for instance. Given that possible world semantics is used to encode these rules, the probability of two rules occurring simultaneously is given by the product of their probabilities. For instance, the probability that a mass has both an oval and round shape is equal to $0.05 \times 0.50 = 0.025$.

```
0.05::feature_shape(Case) :-
  mass(Case, Mass),
  mass_shape(Mass, oval).

0.50::feature_shape(Case) :-
  mass(Case, Mass),
  mass_shape(Mass, round).

0.50::feature_shape(Case) :-
  mass(Case, Mass),
  mass_shape(Mass, irregular).
```

Fig. 1: Probabilistic information from the literature regarding mass shape

Similarly, Fig. 2 also encodes independent rules, each for a characteristic of the mass margin. In this case it becomes obvious that both the microlobulated and spiculated margins have a high correlation with malignancy in the literature, given their high probability of malignancy (70% and 90% respectively).

Figure 3 differs from Fig. 1 and Fig. 2 in that it encodes three mutually exclusive possibilities for the mass density: low, equal, or high (note the new operator ";" for disjunction). The probability of malignancy from the literature is encoded in the top three lines, which can be read as *IF the density of Mass is low, the probability of malignancy is 5%; ELSE IF the density of the Mass is equal, the probability of malignancy is 10%; ELSE IF the density of the Mass is high, the probability of malignancy is 50%*. The density rule is then constructed based on the mutual exclusivity introduced by the `density/1` fact above.

PILP models produce classifiers which are composed by a set of FOL rules, learnt automatically from the data, that represent a disjunctive explanation to the target predicate being learned. Figure 4 presents an example of a PILP model for the target predicate `is_malignant/1`, which explains malignancy in terms of margin OR mass shape and density. Since the rules in this explanation are

```
0.02::feature_margin(Case) :-
  mass(Case, Mass),
  mass_margin(Mass, circumscribed).

0.20::feature_margin(Case) :-
  mass(Case, Mass),
  mass_margin(Mass, indistinct).

0.70::feature_margin(Case) :-
  mass(Case, Mass),
  mass_margin(Mass, microlobulated).

0.90::feature_margin(Case) :-
  mass(Case, Mass),
  mass_margin(Mass, spiculated).
```

Fig. 2: Probabilistic information from the literature regarding mass margin

```
0.05::density(low);
0.10::density(equal);
0.50::density(high).

feature_density(Case) :-
  mass(Case, Mass),
  mass_density(Mass, MassDensity),
  density(MassDensity).
```

Fig. 3: Probabilistic information from the literature regarding mass density

composed of probabilistic literals (`feature_margin/1`, `feature_shape/1`, and `feature_density/1`), the target predicate `is_malignant/1` will also predict a probabilistic value ranging from 0 to 1, even though this is not made explicit in the PILP model. This probability output is computed using the *possible world semantics* [16], and it takes into account the mutual dependency between all the probabilistic literals in the model.

The experiment presented in this work aims at demonstrating that it is possible to use the probabilistic data to build a model that not only obtains good predictive accuracy, but also presents a human-interpretable explanation of the factors that affect the system in study. This model is learnt automatically from the data. In the medical domain it is crucial to represent data in a way that experts can understand and reason about, and as such ILP can successfully be used to produce such models. Furthermore, PILP allows for incorporating in the PBK the confidence of physicians in observations and known values from the literature.

```
is_malignant(Case) :-
  feature_margin(Case).
is_malignant(Case) :-
  feature_shape(Case),
  feature_density(Case).
```

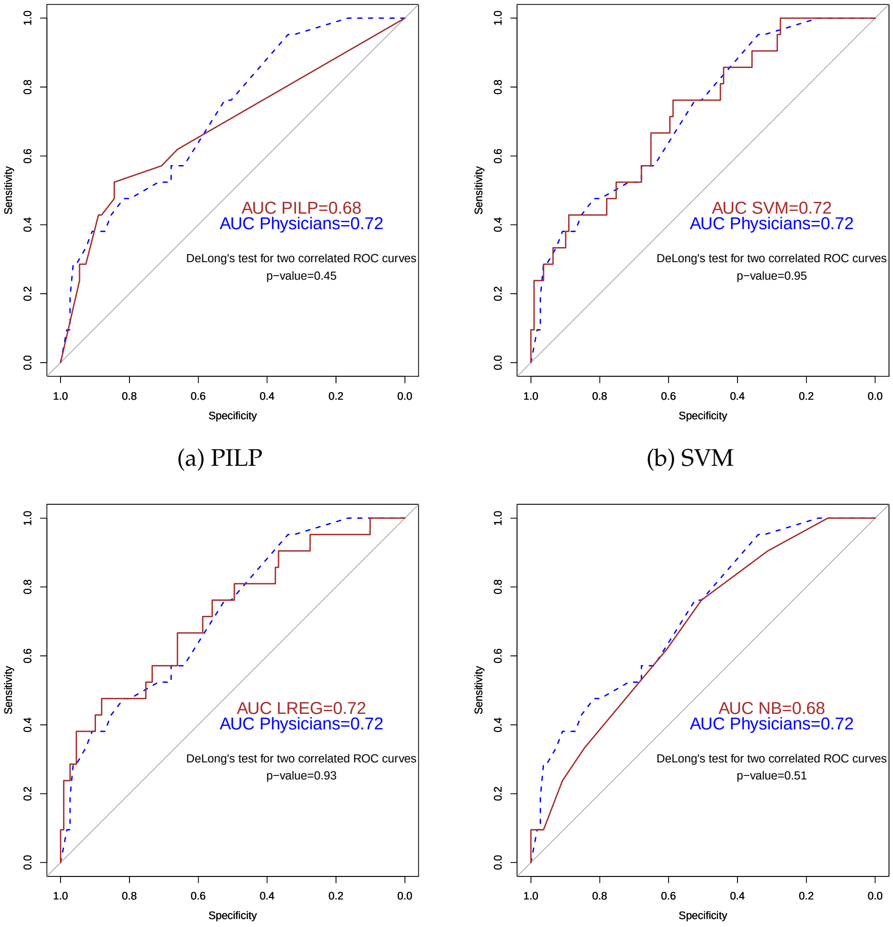
Fig. 4: A PILP model for the target predicate `is_malignant/1`

4 Experiments

The PILP SkILL system [6] was used for these experiments. It runs on top of the Yap Prolog system [7] and uses TopLog [20] as the basis rules generator and the ProbLog Yap library as its probabilistic inference engine. This system was selected because it can perform exhaustive search over the theory search space. Since this is a small dataset, exhaustive search is possible. However, if the dataset were larger there might be scalability issues in using exhaustive search, and so either SkILL with pruning strategies [5] or another PILP system whose search engine is greedy could be used instead (such as ProbFOIL+ [10] or SLIPCOVER [1]). In this experiment, 130 train and tune sets were used to perform leave-one-out cross validation on the dataset, and the predicted values for the test examples were recorded.

In addition to the PILP model described earlier, three other methods were used to compare against PILP in terms of predictive accuracy, using default parameters: a Support Vector Machine (SVM), a Linear Regression (LREG), and a Naive Bayes classifier (NB). The *scikit-learn* python library [21] was used to perform the preprocessing of these experiments for the three non-relational methods. Since these data contain several categorical features, it was necessary to transform them into numerical features to be able to apply these methods. As such, each possible label was first encoded as an integer. Once this was done, each feature was transformed in several auxiliary features, each one of them binary and regarding only one of the labels. This methodology was used to prevent the integer values corresponding to the labels of a feature from being interpreted as being ordered, which would not represent the independence between the labels accurately. Once these operations were performed over all categorical features, a scaler (standardization) was applied so as to reduce all features to mean 0 and unit variance. The predictions for each method were then obtained.

Figure 5 presents the ROC curves for the malignant class and four methods tested: PILP, SVM, LREG and NB. Each sub-figure shows the ROC of the physicians' predictions (blue dashed line) and the ROC of a method (brown solid line), both against the ground truth (confirmed malignancy or benignity of a tumour after excision). Each figure also presents the respective AUCs and the p-value found using DeLong's test for comparing both curves plotted.



(a) PILP

(b) SVM

(c) LREG

(d) NB

Fig. 5: ROC curves, AUCs and p-values for PILP, SVM, LREG and NB methods

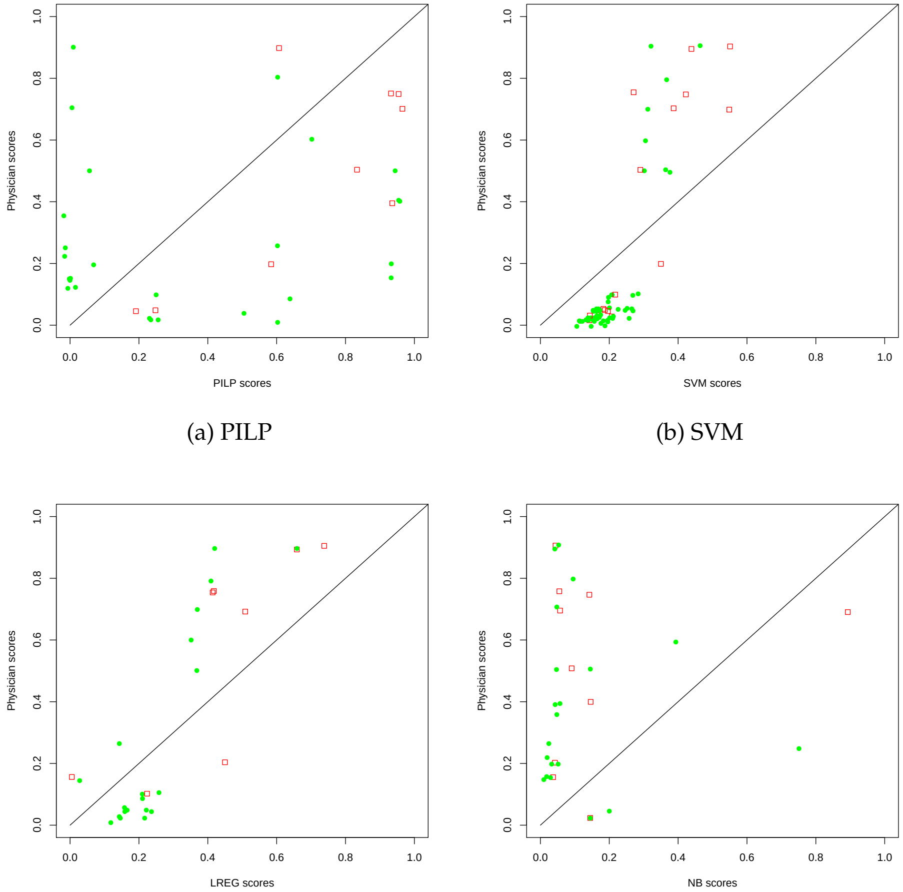
The ROC curves presented in Fig. 5 were compared using DeLong's test for two correlated ROC curves and the difference between them was found to be statistically not significant, thus implying that all methods are statistically indistinguishable from a physician when predicting the degree of malignancy of a patient in this dataset. This experiment established that both PILP and other non-relational methods can successfully mimic the mental model of physicians in what concerns the probabilities of each case in this dataset.

Next, the absolute error of the predictions was analysed. The absolute error is calculated by finding the absolute value of the difference between the prediction and the physicians' score, for a given case. It is relevant to consider the absolute error of predictions because these are the points where the classifiers' predictions disagree with the physicians' mental model, and more information about the performance of the classifier can be obtained from them. Figure 6 shows a plot of the classifiers prediction values (x-axis) against the physicians' prediction values (y-axis), for points where the absolute error was greater than 10%. Points in green (round markers) are cases where the tumour was found to be benign after excision, and conversely points in red (square markers) are cases where the tumour was found to be malignant.

Ideally, malignant prediction by both physician and the classifier should agree and appear on the top right of the plot. Conversely, benign predictions would appear on the bottom left. Points that are plotted below the diagonal line have higher classifier scores than physician scores, and conversely points which are plotted above the diagonal line have higher physician scores than classifier scores.

From the plots in Fig. 6, it is clear to see that the PILP classifier assigns higher malignancy values than physicians do to the confirmed malignancy cases (red points under the diagonal line). This is the case for 8 of the 9 malignant cases, and in the single case where this does not happen, PILP still predicts a reasonably high probability of malignancy (60%). Furthermore, for a malignancy threshold of 0.8, PILP still classifies five malignant cases correctly, whilst this only happens for one case using the physicians' scores. When PILP is compared to the other methods tested, it becomes clear that, in most cases, the other methods do not assign higher scores to malignant points than physicians do (few red points beneath the diagonal line), therefore not being of as much use to physicians as PILP, to aid in the diagnosis of malignant tumours. The ability to identify malignant cases is desirable in medical data since a false negative corresponds to assigning a benign label to a patient who in fact has a malignant tumour.

Since the aim of decision support systems is to aid the process of medical diagnoses, two more models were built based on the results obtained previously. These two models are human and machine models, meaning that they take into account both the physicians' and the classifiers scores. The PILP classifier was selected since it proved to be best at identifying malignant cases that the physicians had difficulty with (unlike other methods). For this reason, two models were analysed: calculating the average of physician and the PILP scores, and calculating the maximum of the physician and the PILP scores. Figure 7



(a) PILP

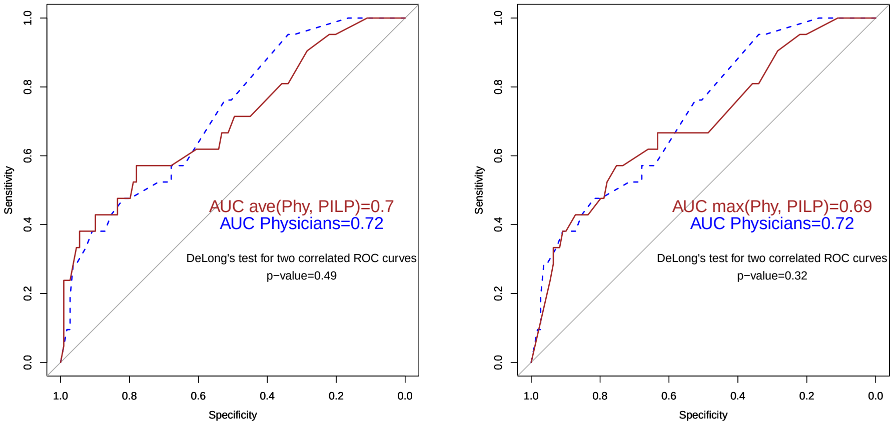
(b) SVM

(c) LREG

(d) NB

Fig. 6: Plot of benign and malignant cases for the PILP, SVM, LREG and NB methods, for errors greater than 0.1, using a negligible amount of jittering

presents the ROCs, AUCs and p-values using DeLong's test for both these models.



(a) Average model

(b) Maximum model

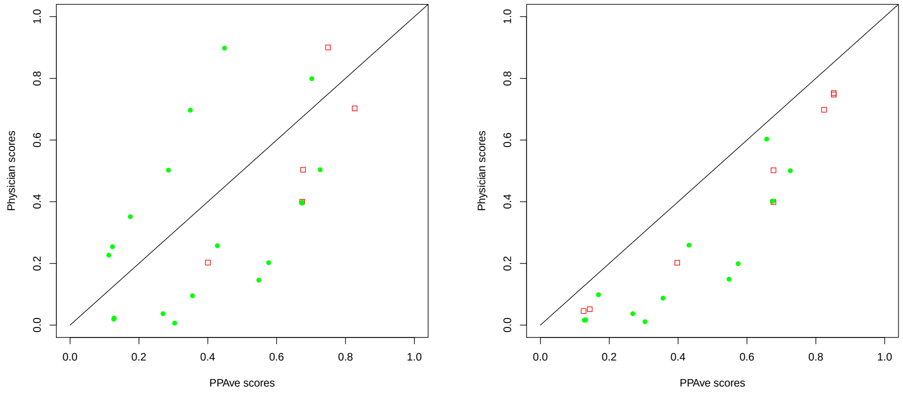
Fig. 7: ROC curves, AUCs and p-values for the average of physician and PILP scores and for the maximum of physicians and PILP scores

The ROC curves plotted in Fig. 7 show no significant difference to the physicians predictive power, similarly to all other classifiers tested. Figure 8 performs the absolute error analysis, plotting the points where these models' predictions and physician's predictions differ by a value greater than 10%.

The scatter plots in Fig. 8 show that the maximum model can now predict higher scores for all malignant points (all red points below the diagonal line). This is to be expected since the model's scores are in effect the maximum score of the PILP and the Physician's model. However, both these models predict higher values for the benign cases as well, which is particularly evident in the case of the maximum model, where there are no points above the diagonal line. Whilst a high recall is a desirable feature in a medical decision support system, the ability to discriminate between malignant and benign cases is also important. The PILP model performs better in this area (Figure 6), since there is a vertical cluster of benign points which are clearly identified by the PILP model as being benign (score of 0.1 or less), and which are no longer present in the combined models analysed here.

Next, the full dataset was used to extract non-trivial knowledge regarding the physician's mental model that is being mimicked and the final theories found are reported in Fig. 9.

From the rules shown in Fig. 9, the first one contains a probabilistic fact related to one mammography descriptor: the shape of a mass. In medical literature, irregular shapes or spiculated margins indicate higher risk of malignancy.



(a) Average model

(b) Maximum model

Fig. 8: Plot of benign and malignant cases for the average and maximum of physician and PILP models, for errors greater than 0.1, using a negligible amount of jittering

This is captured by the system, as well as other features such as no observed increase in mass size and an ultrasound core needle biopsy type. Similarly, the other two rules present features that are evidence of higher risk of malignancy, such as asymmetry, the gauge of the needle and a possible displacement of the needle (offset) during biopsy which can contribute as a confounding factor.

5 Related Work

Relational learning in the form of ILP (without probabilities) has been successfully used in the field of breast cancer. Burnside et al. [8] uncovered rules that showed high breast mass density as an important adjunct predictor of malignancy in mammograms. Later, using a similar dataset, Woods et al. validated these findings [22] performing cross-validation. In another work, Davis et al. used SAYU, an ILP system that could evaluate rules according to their score in a Bayesian network, in order to classify new cases as benign or malignant. Results for a dataset of around 65,000 mammograms consisting of malignant and benign cases showed ROC areas slightly above 70% for Recall values greater than 50% [9]. Dutra et al. showed that the integration of physician's knowledge in the ILP learning process yielded better results than building models using only raw data [13]. The model we use in this paper was presented in more detail in [6] and [5]. One of the datasets used in those works is the same used in this paper, but only for comparing system's execution times. To the best of our knowledge, this is the first work that applies PILP to the area of breast cancer,

```
is_malignant(Case):-
        biopsyProcedure(Case,usCore),
        changes_Sizeinc(Case,missing),
        feature_shape(Case).
is_malignant(Case):-
        assoFinding(Case,asymmetry),
        breastDensity(Case,scatteredFDensities),
        vacuumAssisted(Case,yes).
is_malignant(Case):-
        needleGauge(Case,9),
        offset(Case,14),
        vacuumAssisted(Case,yes).
```

Fig. 9: Theory extracted for physician's mental models.

and illustrates how a probabilistic knowledge representation can be linked with a logic representation to learn stronger and more expressive data models.

6 Conclusion

This work presented a study conducted over breast cancer data, where a PILP model is learnt from the data. This and other machine learning techniques were used to perform a reasonably accurate estimate of breast cancer risk after image-guided breast biopsy, thus alleviating biopsy sampling error. The PILP model combines first order logic with probabilistic data in order to obtain interpretable models that predict probabilities for each new case. The results show that a PILP model can achieve similar results to other traditional classifiers and that its predictions on the test sets are quite close to the experts' predictions. Furthermore, the cases where the models and physicians disagree were analysed in greater detail and it was found that the PILP model consistently assigns high malignancy probabilities to malignant cases, unlike the other models tested. Moreover, the PILP model can explicitly explain why some probability is given to a particular case (using the FOL rules generated), unlike non-relational models. Future work includes studying how changing PILP parameters affects the performance of the system on this and other datasets, as well as studying whether other relevant facts and rules from medical literature can be incorporated in the model.

7 Acknowledgements

This work was partially funded by the North Portugal Regional Operational Programme (NORTE 2020), under the PORTUGAL 2020 Partnership Agreement, and through the European Regional Development Fund (ERDF) as part of project NanoSTIMA (NORTE-01-0145-FEDER-000016). Joana Côrte-Real was funded by the FCT grant SFRH/BD/52235/2013. The authors would like to thank Dr. Elizabeth Burnside for making the dataset used in this paper available to us.

References

1. E. Bellodi and F. Riguzzi. Structure learning of probabilistic logic programs by searching the clause space. *Theory and Practice of Logic Programming*, 15(02):169–212, 2015.
2. W. A. Berg, R. H. Hruban, D. Kumar, H. R. Singh, R. F. Brem, and O. M. Gatewood. Lessons from mammographic histopathologic correlation of large-core needle breast biopsy. *Radiographics*, 16(5):1111–1130, 1996.
3. B. Brancato, E. Crocetti, S. Bianchi, S. Catarzi, G. G. Risso, P. Bulgaresi, F. Piscioli, M. Scialpi, S. Ciatto, and N. Houssami. Accuracy of needle biopsy of breast lesions visible on ultrasound: audit of fine needle versus core needle biopsy in 3233 consecutive samplings with ascertained outcomes. *Breast*, 21(4):449–454, 2012.
4. F. Burbank. Stereotactic breast biopsy: comparison of 14- and 11-gauge mammotome probe performance and complication rates. *Am Surg.*, 63(11):988–995, 1997.
5. J. Côrte-Real, I. Dutra, and R. Rocha. Estimation-Based Search Space Traversal in PILP Environments. In A. Russo and J. Cussens, editors, *Proceedings of the 26th International Conference on Inductive Logic Programming (ILP 2016)*, LNAI, pages –, London, UK, September 2016. Springer. Published in 2017.
6. J. Côrte-Real, T. Mantadelis, I. Dutra, R. Rocha, and E. Burnside. SkILL - a Stochastic Inductive Logic Learner. In *International Conference on Machine Learning and Applications*, Miami, Florida, USA, December 2015.
7. V. Santos Costa, R. Rocha, and L. Damas. The YAP Prolog System. *Journal of Theory and Practice of Logic Programming*, 12(1 & 2):5–34, 2012.
8. J. Davis, E. S. Burnside, I. C. Dutra, D. Page, and V. Santos Costa. Knowledge discovery from structured mammography reports using inductive logic programming. In *American Medical Informatics Association 2005 Annual Symposium*, pages 86–100, 2005.
9. J. Davis, E. S. Burnside, I. C. Dutra, D. Page, R. Ramakrishnan, V. Santos Costa, and J. W. Shavlik. View learning for statistical relational learning: With an application to mammography. In Leslie Pack Kaelbling and Alessandro Saffiotti, editors, *IJCAI-05, Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence, Edinburgh, Scotland, UK, July 30-August 5, 2005*, pages 677–683. Professional Book Center, 2005.
10. L. De Raedt, A. Dries, I. Thon, G. Van den Broeck, and M. Verbeke. Inducing Probabilistic Relational Rules from Probabilistic Examples. In *International Joint Conference on Artificial Intelligence*, pages 1835–1843. AAAI Press, 2015.
11. L. De Raedt and A. Kimmig. Probabilistic (logic) programming concepts. *Machine Learning*, 100(1):5–47, 2015.
12. D'Orsi, C. J. and Bassett, L. W. and Berg, W. A. and et al. *BI-RADS®: Mammography*. American College of Radiology, Inc., 4th edition, 2003. Reston, VA.
13. I. Dutra, H. Nassif, D. Page, et al. Integrating machine learning and physician knowledge to improve the accuracy of breast biopsy. In *AMIA Annual Symposium Proceedings*, pages 349–355, Washington, DC, 2011.
14. A. V. Gonçalves, L. C. Thuler, F. P. Kestelman, P. A. Carmo, C. F. Lima, and R. Cipolotti. Underestimation of malignancy of core needle biopsy for nonpalpable breast lesions. *Rev Bras Ginecol Obstet.*, 33(7):123–131, 2011.
15. J. Halpern. An Analysis of First-Order Logics of Probability. *Artificial intelligence*, 46(3):311–350, 1990.
16. A. Kimmig, B. Demoen, L. De Raedt, V. Santos Costa, and R. Rocha. On the Implementation of the Probabilistic Logic Programming Language ProbLog. *Theory and Practice of Logic Programming*, 11(2 & 3):235–262, 2011.

17. L. Liberman. Percutaneous imaging-guided core breast biopsy: state of the art at the millennium. *Am J Roentgenol.*, 174(5):1191–1199, 2000.
18. L. Liberman, M. Drotman, E. A. Morris, et al. Imaging-histologic discordance at percutaneous breast biopsy. *Cancer*, 89(12):2538–2546, 2000.
19. S. Muggleton and L. De Raedt. Inductive Logic Programming: Theory and Methods. *Journal of Logic Programming*, 19/20:629–679, 1994.
20. S. Muggleton, J. Santos, C. Almeida, and A. Tamaddoni-Nezhad. TopLog: ILP Using a Logic Program Declarative Bias. In *International Conference on Logic Programming*, pages 687–692. Springer, 2008.
21. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
22. R. Woods, L. Oliphant, K. Shinki, D. Page, J. Shavlik, and E. Burnside. Validation of results from knowledge discovery: Mass density as a predictor of breast cancer. *J Digit Imaging*, pages 418–419, 2009.