Probabilistic Logic Models and Their Application to Breast Cancer

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 to 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. Experiments show that the probabilistic model produced can not only match the predictions of a team of experts in the area, but also produce meaningful rules which output better calibrated probability values.

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) [11] 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 [9], 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).

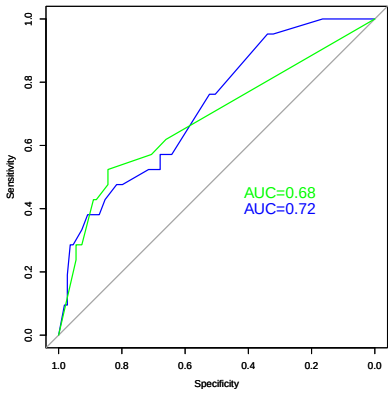
Breast cancer is one of the most common forms of cancer and mammograms are the most commonly used technique to detect patients at risk. Image-guided core needle biopsy of the breast is then performed to decide on surgery. Biopsy is a necessary, but also aggressive, high-stakes procedure. The assessment of malignancy risk following breast core biopsy is imperfect and biopsies can be *non-definitive* in 5-15% of cases [1]. 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) [7] 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 be relevant to help clinicians assessing new cases.

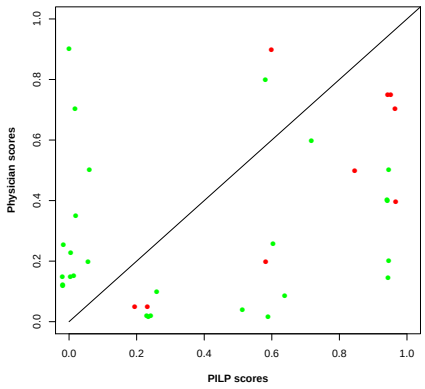
2 Methodology and Results

The dataset used for this experiment contains 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) was used as the target



(a) ROC curves (and area under the curve) of PILP and Physicians, both against ground malignancy after excision



(b) Plot of benign and malignant cases for errors greater than 0.1, using a negligible amount of jittering

Fig. 1: PILP and physician ROC curves for all cases (left) and PILP and physician scores for large errors

predicate. This value was assigned by a multidisciplinary group of physicians analysing that case. Thus, the target probabilities of examples represent the perceived chance of malignancy for each patient.

Regarding the domain knowledge incorporated in the PBK, breast cancer literature values were used to complement the information on the characteristics of masses, since these values are relied on by physicians when they 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 and density (all part of the BIRADS terms).

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. 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.

In this experiment, 130 train and tune sets conserving the positive/negative ratio of the full dataset were used to perform leave-one-out cross validation on the dataset, and the predicted values for the test examples were recorded. Figure 1a presents the ROC curves for the malignant class and the area under the curve (AUC) for PILP's predicted test values (green) and for the physicians orig-

```
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. 2: Theories extracted for physician's mental models.

inal predictions (blue), both against the ground truth (confirmed malignancy or benignity of a tumour after excision).

The ROC curves presented in Fig. 1a were compared using DeLong's test for two correlated ROC curves and its p-value was found to be 0.4476, thus implying PILP's classifier and a physician are statistically indistinguishable when predicting the degree of malignancy of a patient in this dataset. This experiment established that PILP 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 PILP predictions was analysed. The absolute error is calculated by finding the absolute value of the difference between the PILP prediction and the physicians' score, for a given case. In 94 cases (72%), the PILP prediction lies within at most 0.1 of the physician's value. For the remaining 36 cases (28%), the PILP and physicians' values were compared. Figure 1b shows a plot of the PILP prediction value (x-axis) against the physicians' prediction value (y-axis). Points in green are cases where the tumour was found to be benign after excision, and conversely points in red are cases where the tumour was found to be malignant. This plot shows that for 8 of the 9 malignant cases, PILP predicts a significantly higher malignancy value than physicians do (red points under the diagonal line). 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. This behaviour 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.

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. 2. From the rules shown in Fig. 2, the first one contains a probabilistic fact related to one mammography descriptor: the shape of a mass. In medical literature, irregular or spiculated shapes indicate higher risk of malignancy. 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.

This experiment can also be compared against the predictions (using the same data) of a Naive Bayes classifier using a similar methodology as discussed by Kuusisto *et al.* [10], and it was found that the probabilities produced using PILP are much closer to the values given by the physicians than the probability values produced by the Naive Bayes classifier, making PILP predictions much closer to the actual values that the physicians use to asses their patients.

3 Related Work

Relational learning in the form of ILP (without probabilities) has been successfully used in the field of breast cancer. Burnside et al. [4] 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 [12] 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% [5]. 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 [8]. The model we use in this paper was presented in more detail in [3] and [2]. 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, and illustrates how a probabilistic knowledge representation can be linked with a logic representation to learn stronger and more expressive data models.

4 Conclusion

This work presented a machine learning technique that can perform a reasonably accurate estimate of breast cancer risk after image-guided breast biopsy, thus alleviating biopsy sampling error. This 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, in the cases where PILP predictions are significantly different from expert values, PILP consistently assigns high malignancy probabilities to malignant cases. Moreover, this model can explicitly explain why some probability is given to a particular case (using the FOL rules generated), unlike non-relational models. These results are

encouraging, but still there is room for improvements. 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.

References

1. 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.
2. Joana Côrte-Real, Inês Dutra, and Ricardo Rocha. Estimation-based search space traversal in pilp environments. In *26th International Conference on Inductive Logic Programming*, London, UK, 2016.
3. Joana Côrte-Real, Theofrastos Mantadelis, Inês Dutra, Ricardo Rocha, and Elizabeth Burnside. Skill - a stochastic inductive logic learner. In *14th IEEE International Conference on Machine Learning and Applications (ICMLA 2015)*, Miami, FL, USA, December 2015. IEEE, IEEE.
4. 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.
5. 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.
6. L. De Raedt and K. Kersting. Probabilistic inductive logic programming. In *International Conference on Algorithmic Learning Theory*, pages 19–36. Springer, 2004.
7. 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.
8. 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.
9. J. Halpern. An Analysis of First-Order Logics of Probability. *Artificial intelligence*, 46(3):311–350, 1990.
10. F. Kuusisto, I. Dutra, H. Nassif, Y. Wu, M. Klein, H. Neuman, J. Shavlik, and E. Burnside. Using Machine Learning to Identify Benign Cases with Non-Definitive Biopsy. In *International Conference on e-Health Networking, Application & Services*, page 283–285. IEEE, 2013.
11. S. Muggleton and L. De Raedt. Inductive Logic Programming: Theory and Methods. *Journal of Logic Programming*, 19/20:629–679, 1994.
12. 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.