Integrating Machine Learning and Physician Knowledge to Improve the Accuracy of Breast Biopsy

I. Dutra[1], H. Nassif[2], D. Page[2], J. Shavlik[2], R. M. Strigel, MD, MS[2], Y. Wu[2], M. E. Elezaby, MD[2], E. Burnside, MD, MPH, MS[2]
[1] University of Porto, Porto, Portugal; [2]University of Wisconsin, Madison, USA

Abstract

In this work we show that combining physician rules and machine learned rules may improve the performance of a classifier that predicts whether a breast cancer is missed on percutaneous, image-guided breast core needle biopsy (subsequently referred to as "breast core biopsy"). Specifically, we show how advice in the form of logical rules, derived by a sub-specialty, i.e. fellowship trained breast radiologists (subsequently referred to as "our physicians") can guide the search in an inductive logic programming system, and improve the performance of a learned classifier. Our dataset of 890 consecutive benign breast core biopsy results along with corresponding mammographic findings contains 94 cases that were deemed non-definitive by a multidisciplinary panel of physicians, from which 15 were upgraded to malignant disease at surgery. Our goal is to predict upgrade prospectively and avoid surgery in women who do not have breast cancer. Our results, some of which trended toward significance, show evidence that inductive logic programming may produce better results for this task than traditional propositional algorithms with default parameters. Moreover, we show that adding knowledge from our physicians into the learning process may improve the performance of the learned classifier trained only on data.

1 Introduction

Image-guided percutaneous core needle biopsy is the standard of care for the diagnosis of a suspicious finding in the breast. Unfortunately, the assessment of malignancy risk following breast core biopsy is imperfect and biopsies can be "non-definitive" in 5-15% of cases[1]. This category includes discordant biopsies, high risk lesions (atypical ductal hyperplasia, lobular carcinoma in situ, radial scar, etc.), and insufficient sampling. A non-definitive result means the chance of malignancy remains high due to possible sampling error (i.e. obtained biopsy is not representative of the suspicious finding), for which surgical excisional biopsy is performed[2]. Image-guided breast core biopsies may therefore result in missed breast cancers (false negatives) or unnecessary surgery (false positives). In the US, the 112 million women over the age of 20 years[3] 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 2010. As a result of "non-definitive" biopsies, approximately 35,000–105,000 of these women will require additional biopsies secondary to judged inadequacy of breast core biopsy. In this work we show how knowledge-based clinical decision support systems, with the help of physicians, may solve this problem.

Our goal is to investigate if adding specialist knowledge to machine learned rules can improve a final classifier prediction of breast cancer. Most of the previous research investigating machine learning techniques for breast cancer risk prediction used conventional algorithms like artificial neural networks or Bayesian networks that use either expert opinion or are trained on data, not a combination.[7,8,9,10,11] An experiment like this, where expert knowledge is added to help improve the performance of the classifier, was performed by Lavrac and Dzeroski with the LINUS system, to learn rules for early diagnosis of rheumatic diseases[12]. In their experiments, in addition to the attribute-value description of patient data, LINUS was also given background knowledge provided by the medical specialist in the form of typical co-occurrences of symptoms. Our study is different for three reasons. First, we investigate how to merge machine learning from data and physician expertise to optimize a decision support tool. Second, we use first order rules with variables instead of only adding ground facts to the background knowledge, as was done in the experiment using LINUS. Third, we focus on the previously unexplored application area: the problem of classifying "non-definitive" biopsies. Therefore, our final goal is to develop a classifier that integrates specialist knowledge and machine learning, in order to produce classifiers that are better than either alone.

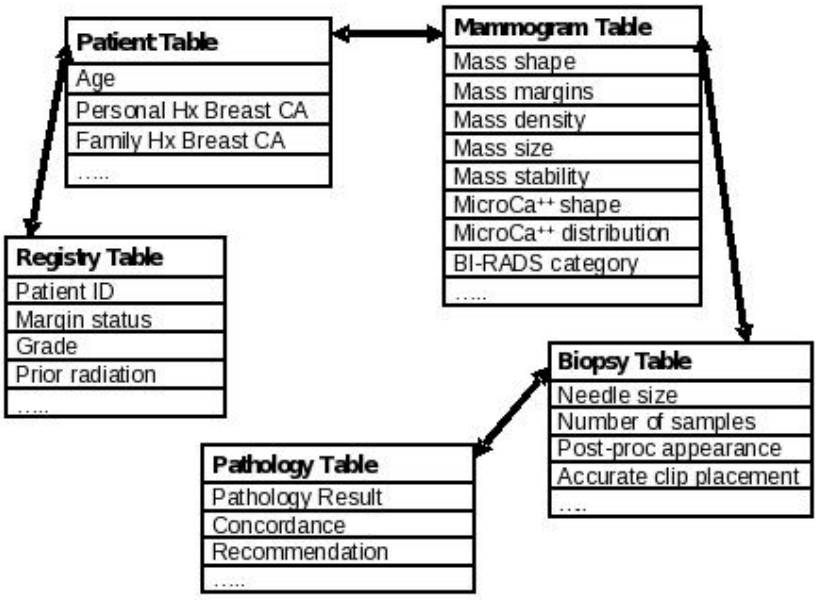

Figure 1: Multi-relational tables necessary for estimating risk of breast cancer following core biopsy.

2 Materials and Methods

Our study included 1228 consecutive image-guided core biopsies, of which 890 were benign and 338 malignant. All 1228 biopsies were scrutinized at our concordance conference and out of the 890 benign biopsies, 94 were deemed non-definitive. Surgical excision was ultimately performed for these 94 cases and 15 of which were upgraded (i.e. found to be malignant). Breast core biopsies were classified as concordant (i.e. the pathologic result explained the imaging findings), discordant (i.e. the pathologic diagnosis did not explain the imaging findings), or 'ars' for results of atypia (including atypical ductal hyperplasia, atypical lobular hyperplasia, lobular carcinoma in situ) and radial scar. The non-definitive cases in this study included the discordant cases and 'ars' cases. Concordance at our institution is determined by the radiologist performing the core biopsy in conjunction with a weekly radiologic-pathologic consensus conference attended by multiple radiologists and pathologists as described in Salkowski *et al*[2]. The multiple radiologists contributing to consensus have a range of backgrounds and experience, including breast fellowship training at multiple different institutions.

biopsy procedure	abn. disappeared (1/0)
pathology priority	mass present (1/0)
pathology description	asymmetry
concordance	architectural distortion
mammography birads	calcs (1/0)
ultrasounds birads	mass shape
MRI birads	mass margins
combined birads	mass density
biopsy needle type	calcifications
biopsy needle gauge	calcification distribution
clip in site of biopsy (1/0)	mass size

Table 1: Attributes used in this study.

The main attributes used in our study are spread out through five tables as shown in Figure 1, but for the experiments in this paper, we filtered and combined these into one large table, with one record per biopsy, containing only the attributes considered relevant for our focused study (e.g., free-text attributes were not considered in our study). These attributes are presented in Table 1. Most of them are self-explanatory. The attribute "concordance" can have values 'd' for discordant, 'ars' for atypical and radial scar, and 'i' for insufficient information.

Two of our physicians (Dr. Burnside and Dr. Elezaby) created rules such as the ones shown in Figure 2 related to the non-definitive and upgraded cases. Note that these rules are as simple as possible, not requiring that the specialist thoroughly inspects all cases. We left the exhaustive task of inspecting every single case to the machine learning system. We believe that even using simple specialist's rules, the knowledge encoded there will be sufficient to produce a classifier that performs better than one that is totally blind to any specialist's knowledge.

In our experiments, we used the Aleph system, developed by Ashwin Srinivasan, at the University of Oxford, England[13] and the WEKA system, developed at University of Waikato, New Zealand[14]. Aleph is a machine learning (ML) system within the area of ML known as Inductive Logic Programming (ILP). Aleph takes as input positive and negative observations, as well as relevant background knowledge (a description of the observations). In our study, a positive observation is a biopsy initially considered benign at core biopsy and subsequently proved to be malignant at surgical excision (called an upgrade). A negative example is a biopsy deemed benign that is confirmed as benign by follow-up by our cancer registry match. Each observation in the background knowledge is described through first order logic facts. E.g., the fact `needlegauge(25,9)` represents the needlegauge used for observation 25. Unlike WEKA, Aleph naturally handles first order rules that are automatically extracted from the background knowledge. Aleph produces as output one or more rules that, together with the background knowledge, provide the appropriate observation labels. Therefore, the output of Aleph would be something like the rule showed in Figure 3.

(1) Probability of upgrade increases for stereotactic biopsy if the finding is microcalcifications and biopsy results reveals atypical ductal hyperplasia (ADH)
upgrade(A) IF
  pathDx(A,atypical_ductal_hyperplasia),
  calcifications(A,'a,f') and
  biopsyProcedure(A, stereo)

(2) Probability of Upgrade increases for ultrasound (US) biopsy if the finding is a mass and biopsy results reveals benign breast tissue
upgrade(A) IF
  concordance(A,d),
  biopsyProcedure(A, US_core) and
  pathDx(A,benign_breast_tissue)

Figure 2: Physicians rules coded in logic. Arguments starting with a capital letter are logical variables that are bound to specific patient data. Space limitations preclude full explanations of the rules, but we used lengthy predicate and variable names to partially document them.

This rule has the following English translation: biopsy A is an upgrade if the case is discordant (concordance(A,d)), the needle gauge used in the biopsy is 9 and the abnormality did not show in the biopsy. A case is deemed discordant when the pathologic diagnosis does not explain the imaging findings, as determined during an Imaging/Pathology consensus conference[2].

```
upgrade(A) if
        concordance(A,d) and
        needlegauge(A,9) and
        abndisappeared(A,1)
```

Figure 3: Example of Aleph rule.

With Aleph, we performed three experiments whose main objective was to prove if the rules given by our physicians helped Aleph search for new rules. In the first experiment, using only the data, we learned ILP rules to predict upgraded cases (hereafter called

Experiment	TP	FP	FN	TN	Recall	Precision	F1
Human rules alone	9	45	6	34	**0.60**	0.16	0.26
ILP rules alone	5	11	10	68	0.33	0.31	0.32
Human + ILP	8	13	7	66	**0.53**	**0.38**	**0.44**
WEKA-TAN	3	9	12	70	0.20	0.25	0.22
WEKA-TAN+Human rules	3	7	12	72	0.20	0.30	0.24

Table 2: Performance of the classifiers.

ILP rules alone). In the second experiment, our physicians formulated a theory in first order logic to explain the reason for the upgrades (hereafter called **Human Rules alone**). We then introduced our physicians rules as part of the background knowledge and re-trained our classifier to predict upgraded cases (this combination is the third experiment and is called **Human + ILP**). In other words, in this latter experiment, we extend our background knowledge of facts with rules provided by our physicians. Aleph will learn new rules from this new background knowledge. For **ILP rules alone** and **Human + ILP**, we ran Aleph, with the same set of parameters, the same 15 folds and the same stratified 15 fold cross-validation technique for training. We created a new function to evaluate each rule in order to take into account the imbalance between the number of positive and negative examples. We set the minimum required accuracy for a generated rule to 0.1, and allowed it to cover up to 1 negative example. We used the default values for the remaining Aleph parameters. To evaluate the results we applied a two-tailed paired t-test to compare both experiments, as implemented in Excel. For **Human rules alone**, we generated a theory (disjunction of the 18 rules provided by our physicians) and applied it to the data.

WEKA is a tool that implements traditional propositional machine learning algorithms such as decision trees and Bayesian networks. Input data in WEKA needs to be in the form of a table of instances and attributes (for example, an Excel table), where, usually, the last attribute is the class variable. Given this table, the label variable, and the chosen algorithm with its parameters, WEKA produces a classifier to predict the class variable.

For the experiments with WEKA, we experimented with several classification algorithms, but report only for the algorithm that produced the best results (a Bayesian network using Tree-Augmented Naive Bayes – TAN[15]). The experiments were performed using the Experimenter module and the default parameters for the learning algorithms. We did set the following parameters equally for each algorithm: statistical significance test (corrected paired t-test with a significance level of 0.01. WEKA has two ways of performing t-test: standard paired and corrected paired t-test. The standard paired t-tester assumes the samples are independent. However, due to the cross validation used, the samples are not independent. Ignoring this assumption generally gives very high type I errors, i.e., the test saying there is a difference between the tested algorithms while, in fact, there is not. The corrected t-test uses a fudge factor to counter the dependence between samples which in practice results in acceptable type I errors), the confidence interval (95%) and the number of times each experiment was executed (= 30). We set two experiments for WEKA. One that used the original dataset (thereafter called **WEKA-TAN**), and another one enriched with 18 new binary attributes, each one corresponding to the coverage of one rule provided by our physicians (thereafter called **WEKA-TAN + Human rules**). For these 18 attributes, a value of zero for an instance means that the rule does not cover the example, a value of 1 means that the rule covers the example.

Our reported results are based on cross-validation. Therefore, they always reflect performance on test data that was not used in learning. When applicable, we measure prediction accuracy using precision-recall curves. Precision gives the rate of correctly classified positives (positive prediction value – PPV) while Recall gives the true positive rate.

3 Results

Table 2 shows all results for Aleph, WEKA and the theory provided by our physicians. For each experiment, we report the contingency table (TP - True Positives, FP - False Positives, FN - False Negatives and TN - True Negatives), Recall, Precision and F1 statistics (the harmonic mean of Recall and Precision). The performance of our physicians rules is shown in the first row of Table 2 (**Human rules alone**). The second row shows the performance of Aleph when

learning only from the data (**ILP rules alone**). The third row shows the performance of the combination between our physicians rules and Aleph (**Human + ILP**). The last two rows show the results of the classifiers produced with WEKA.

The accuracies of both Aleph classifiers are approximately the same, 77%. They mainly differ on the Precision and Recall measures. When combining the human rules with ILP, the resulting classifier boosts recall from .33 (**ILP rules alone**) to .53 (**Human + ILP**) (p=0.08) and precision from .31 to .38 (p=0.35). Note that the F1 measure, the harmonic mean of Precision and Recall, also improves. In other words, our physicians rules succeed in improving recall (sensitivity) with a concomitant increase in precision. We can catch fully half of the false negative biopsies (8 of 15 for Recall > 0.5) at the cost of re-examining just under twice that number of actual negative biopsies (13 of 21 upgrade predictions, for Precision > 0.35). Though neither of these results is statistically significant, the recall difference demonstrates a trend. Recall (sensitivity) is extremely important in breast cancer diagnosis because missing a breast cancer has a high penalty. The Recall obtained with this classifier is nearly the same as the one obtained by our physicians, with this classifier producing a substantial reduction in the number of false positives compared with our physicians. It is possible that a refinement of the rules provided by the specialist could further improve these results.

The Bayesian network generated by WEKA produces better results when classifying negative instances (non-upgrade biopsies), but behaves very poorly when classifying the upgrades (Recall = 0.2 for both classifiers). When adding new attributes to the data (encoded rules provided by our physicians), the Precision improves. These results were obtained with the default threshold 0.5. In Figure 4, we show the Precision-Recall (PR) curve corresponding to the results produced by **WEKA-TAN** and **WEKA-TAN + Human rules** when varying the threshold. We also show how our ILP results and our physicians theory compares to these curves. When we use ILP to produce rules (square symbol in Figure 4), the results are close to both PR curves. When we add our physicians rules to the ILP training, performance improves (triangle symbol in Figure 4). In both versions (**WEKA-TAN** and **WEKA-TAN + Human rules**), with a threshold of 0.5 (the WEKA default), the number of true positives is 3, which is far below what we can achieve with Aleph and with Aleph with human rules. The number of true negatives is 70 for WEKA and 72 for WEKA with added attributes. The Bayesian classifier performs slightly better on predicting true negatives for the 0.5 threshold value. Varying the threshold to 0.0005, we can produce a classifier that achieves Recall of 100% with Precision of approximately 20%. In that situation, the number of women that need to go through an extra examination (false positives) is still reasonable (approximately 60 out of 79), given that we can correctly predict all non-definitive biopsies that are upgraded to malignancy. Note, however, that this result is likely to be optimistic since we have not provided a method for selecting a threshold that promises to produce no (or very few) false negatives on future cases. We obtained the threshold value of 0.0005 by looking at all the predictions on the held-aside cases in our cross-validation experiment.

In Figure 4, we also show how all results (ILP and WEKA) compare with the physicians rules alone (the diamond symbol). The combination of the theory provided by our physicians with ILP improves precision without a decrease in recall. In other words, from a clinical standpoint, we could have saved 19 surgeries without missing a cancer. We believe these results are encouraging and motivate further research.

Just to illustrate the differences between the results of **ILP rules alone** and **Human Rules+ILP**, one of the rules found in the latter experiment, is:

```
upgrade(A) :- binSpecimens(A,gt6), phyRule1_2(A), phyRule1_3(A).
```

which combines one attribute of our original background knowledge with two rules provided by our physicians, with both of them defined as:

```
phyRule1_2(Id) :-
      pathDx(A,adh),
      calcifications(Id,'a,f').
```

```
phyRule1_3(Id) :-
      pathdxabbr(Id,adh),
      calcification_distribution(Id,g).
```

`phyRule1_2(Id)` is a subset of the rule 1 given in Figure 2 ('adh' stands for atypical_ductal_hyperplasia). When learning with only attributes, Aleph never found a rule that combined `binSpecimens` with `calcification`

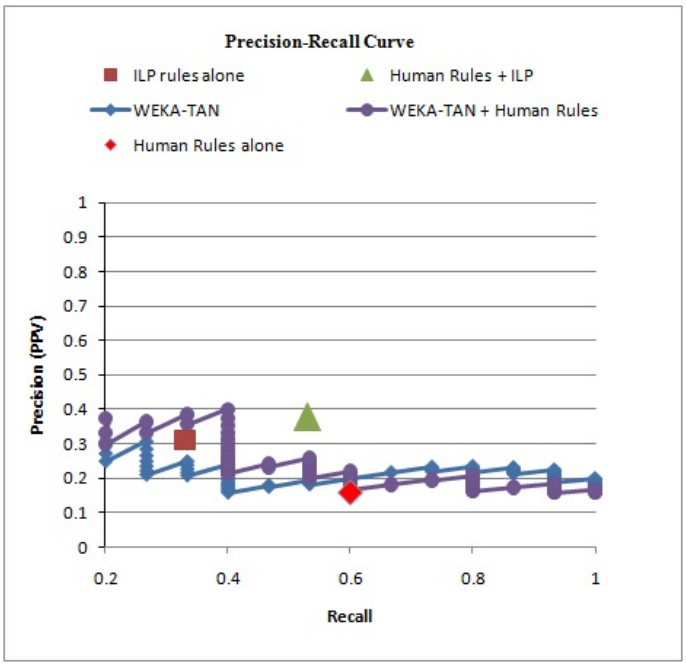

Figure 4: PR curve for WEKA-TAN and WEKA-TAN + Human rules compared with, ILP rules alone, Human Rules+ILP, and Human Rules alone.

`_distribution` **AND** `calcifications`. It also never found a rule that combined `binSpecimens` with `calcification_distribution`. On the other hand, some rules were found that combined `binSpecimens` with `calcifications`. Adding physicians rules to the background knowledge influenced the way Aleph guided the search yielding better results.

Of note, the new rules found by Aleph have the potential to uncover associations predictive of upgrade not yet observed or appreciated by the physicians. We expect to delve more deeply into the rules developed by Aleph to further characterize these associations in future work.

4 Conclusions and Future Work

In this work we showed how advice, in the form of first order rules, given by specialists in breast imaging, may improve the quality of a learned classifier. We focused our study on the difficult task of discriminating non-definitive cases (those benign cases not adequately evaluated by breast core biopsy, which may subsequently upgrade to malignancy). We used a small dataset of non-definitive and upgraded cases to prove our concept. Specialists studied all the biopsies and formulated a theory in first order logic to explain the upgraded cases. We used this theory to help guide the search for new rules. We also compared the performance of the rule-based classifier with a traditional propositional learner based on a Bayesian network. Our results show that (1) the use of advice from specialists may improve the performance of both classifiers (based on rules and on Bayesian networks) and (2) inductive logic programming performs better in doing this task than traditional classifiers. We believe these results are encouraging, and have prompted work in new directions: (1) collection of more data related to non-definitive biopsies and upgrades to further refine the algorithm, (2) construction of an iterative system that can interact with specialists in order to refine rules and use expert advice to continuously improve the quality of the classifiers, (3) application of these techniques to other medical domains, and (4) taking advantage of the full power of the first order logic by taking into account multiple data tables.

References

1. Laura Liberman. Percutaneous imaging-guided core breast biopsy: State of the art at the millennium. *Am. J. Roentgenol.*, 174(5):1191–1199, 2000.

2. Lonie R. Salkowski, Amy M. Fowler, Elizabeth S. Burnside, and Gale A. Sisney. Utility of 6-month follow-up imaging after a concordant benign breast biopsy result. *Radiology*, 258(2):380–387, 2011.

3. National vital statistics report. Technical report, National Center for Health Statistics, 1998.

4. Lisa Torrey, Trevor Walker, Jude Shavlik, and Richard Maclin. Using advice to transfer knowledge acquired in one reinforcement learning task to another. In *In Proceedings of the Sixteenth European Conference on Machine Learning*, pages 412–424, 2005.

5. Richard Maclin, Jude Shavlik, Trevor Walker, and Lisa Torrey. A simple and effective method for incorporating advice into kernel methods. In *Proc of the 21st National Conf on Artificial Intelligence (AAAI'06)*, 2006.

6. Richard Maclin, Edward Wild, Jude Shavlik, Lisa Torrey, and Trevor Walker. Refining rules incorporated into knowledge-based support vector learners via successive linear programming. In *Proceedings of the 22nd national conference on Artificial intelligence - Volume 1*, pages 584–589. AAAI Press, 2007.

7. Y Wu, M L Giger, K Doi, C J Vyborny, R A Schmidt, and C E Metz. Artificial neural networks in mammography: application to decision making in the diagnosis of breast cancer. *Radiology*, 187(1):81–87, 1993.

8. Nick Street, O. L. Mangasarian, and W. H. Wolberg. An inductive learning approach to prognostic prediction. In *Proceedings of the Twelfth International Conference on Machine Learning*, pages 522–530. Morgan Kaufmann, 1995.

9. Hussein A. Abbass. An evolutionary artificial neural networks approach for breast cancer diagnosis. *Artificial Intelligence in Medicine*, 25:265–281, 2002.

10. Turgay Ayer, Oguzhan Alagoz, Jagpreet Chhatwal, Jude W. Shavlik, Charles E. Kahn, and Elizabeth S. Burnside. Breast cancer risk estimation with artificial neural networks revisited. *Cancer*, 116(14):3310–3321, 2010.

11. Jesse Davis, Elizabeth S. Burnside, Inês de Castro Dutra, David Page, and Vítor Santos Costa. Knowledge discovery from structured mammography reports using inductive logic programming. In *American Medical Informatics Association 2005 Annual Symposium*, pages 86–100, 2005.

12. N Lavrac and S. Dzeroski. *Inductive Logic Programming: Techniques and Applications*. Ellis Horwood, New York, 1994. chapter 11.

13. Ashwin Srinivasan. *The Aleph Manual*, 2001.

14. Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The weka data mining software: an update. *SIGKDD Explor. Newsl.*, 11:1018, November 2009.

15. N. Friedman, D. Geiger, and M. Goldszmidt. Bayesian network classifiers. *Machine Learning*, 29:131–163, 1997.