Leveraging Expert Knowledge to Improve Machine-Learned Decision Support Systems

Finn Kuusisto, MS[1], Inês Dutra, PhD[2], Mai Elezaby, MD[1], Eneida A. Mendonça, MD, PhD[1], Jude Shavlik, PhD[1], Elizabeth S. Burnside, MD, MPH, MS[1]
[1] University of Wisconsin, Madison, USA [2]University of Porto, Porto, Portugal

Abstract

While the use of machine learning methods in clinical decision support has great potential for improving patient care, acquiring standardized, complete, and sufficient training data presents a major challenge for methods relying exclusively on machine learning techniques. Domain experts possess knowledge that can address these challenges and guide model development. We present Advice-Based-Learning (ABLe), a framework for incorporating expert clinical knowledge into machine learning models, and show results for an example task: estimating the probability of malignancy following a non-definitive breast core needle biopsy. By applying ABLe to this task, we demonstrate a statistically significant improvement in specificity (24.0% with p=0.004) without missing a single malignancy.

Introduction

Collaborations between medical domain experts (MDE) and computer science experts (CSE) often involve the use of machine learning to develop predictive models aimed at improving patient care. Unfortunately, standardized, complete, and sufficient training data for machine-learning algorithms is rarely available for a variety of reasons including variability of practice between physicians as well as institutions, low disease prevalence on a population level, and confidentiality issues. The difficulty inherent in collecting large, high quality datasets represents a major challenge in the development of machine learned models for decision support. One of the solutions to this challenge is to incorporate the clinical experience and intuition of MDEs, that may help compensate for a lack of large training datasets.[1,2] In fact, some successful cases of integrating expert knowledge with predictive and analytical models are available in the literature.[3,4] As it is nearly impossible for MDEs, who are not programmers, to contribute their expertise directly to the software, we argue that there is a need for a framework that improves close collaboration between MDEs and CSEs to provide a method for shared dialog. Rather than solely providing training through a set of examples, it would be much more valuable if the MDEs could (a) explain what the machine learner is doing wrong and (b) explain how to fix the current problem in a manner that will generalize to similar future cases. This dialog is the basic idea motivating our development of Advice-Based-Learning (ABLe). In ABLe, MDEs provide advice, and the learning algorithm is able to decide how best to absorb it, possibly rejecting the advice or refining it based on the available data. Based on continual observation of model performance, the MDEs can provide additional advice.

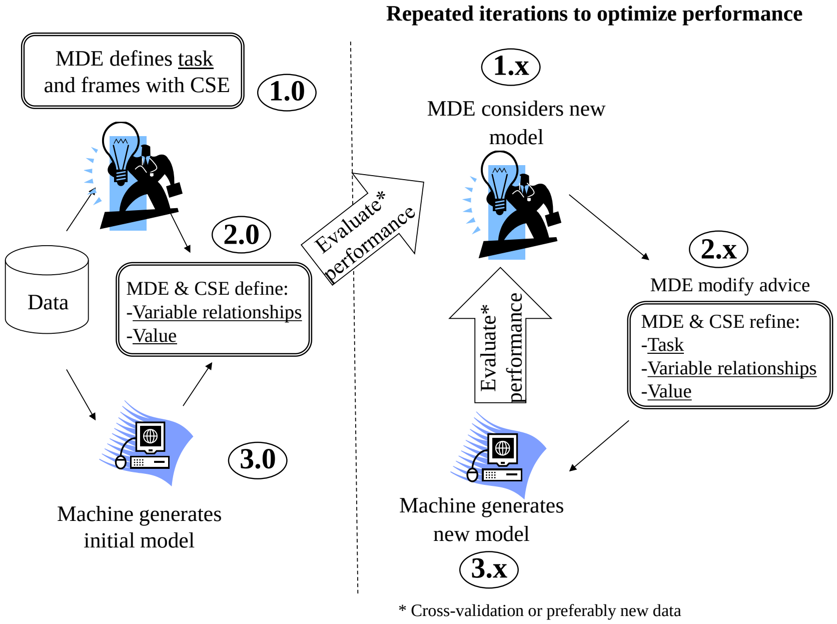

Figure 1: The ABLe Framework.

The ABLe Framework

Our ABLe framework (Figure 1) includes: (1) definitions and (2) iterative steps.

Appears in the *AMIA Joint Summits on Translational Science, Clinical Research Informatics*, March 23-27, 2015, San Francisco, CA

The definitions are:

Task The problem and scope with quantification of appropriate predictive variables.
Variable Relationships Combinations of variables that are particularly important for the task.
Parameter Values Algorithm settings or parameters that best represent the clinical objective.

In the process of developing a decision support system, definitions will be unique to the clinical goal. Modeling techniques should be chosen based on both the data and the task, but this framework can provide a way by which the MDEs and CSEs can interact.

Regarding the iterative steps, we follow a similar process to Gibert et al.[3] In Steps 1 through 3, the MDEs and CSEs interact to establish an initial model. In Step 1, the MDEs (physicians in our example) define a task to address, provide the data, determine what variables will be used from the data available, and determine what is the desired outcome. Any machine learning technique, which is supervised,[2] i.e., can be evaluated against a gold standard (or ground-truth), needs the definition of the variable(s) corresponding to the outcome as determined in the task definition. At this point, the CSEs are involved in picking an appropriate algorithm, and defining an appropriate formal language to represent tasks and variables, which is used to interface with the MDEs.

In Step 2, the physicians and computer scientists interact to produce an initial set of variable relationships and value specifications. The variable relationships correspond with clinician intuition about predicting the chosen task based on relevant knowledge (e.g. the literature) and available data. This advice is encoded in a way that allows it to be incorporated directly into the chosen algorithm. There are multiple ways for prior knowledge to be incorporated into learning algorithms,[2] e.g., using expert-constructed network structures for graphical models[1]. The value specifications correspond to proper selection of algorithm parameters and other experimental settings in order to obtain clinical significance. For example, the physicians can help the computer scientists specify costs of misclassification or a weighting scheme for importance of examples. Finally, in Step 3 the initial model is trained and produces results on an unseen set of data.

Steps 1.x through 3.x show the iterative refinement process that occurs after initial model production. In Step 1.x, the MDEs consider the results produced by the model. Another interaction between MDEs and CSEs takes place, and previous definitions of the task, variable relationships, or value specifications are then modified in Step 2.x. For example, if the current task definition proves to be too challenging, the MDEs can redefine the task in the most fruitful direction. Similarly, the MDEs can modify, remove, or provide additional variable relationships or value specifications. Finally, in Step 3.x, a new model is trained and produces results on unseen data, which again leads to Step 1.x+1. This process continues until the MDEs are satisfied with the results produced by the model learned.

To illustrate this framework, we use it on an example task, predicting undiagnosed malignancy in the setting of a benign but non-definitive image-guided breast core biopsy. In the next two sections, we introduce the task and explain how we used ABLe to achieve our best performance. Although we give a specific example, the same framework would work for a wide range of medical tasks.

Task Prediction

When a patient presents with a suspicious breast lesion, a diagnostic mammogram and possibly ultrasound are performed to further define the abnormality. If the finding remains suspicious, a core needle biopsy (CNB) is often recommended. In this procedure, a needle is inserted into the breast under imaging guidance to remove small samples ("cores") of the targeted breast abnormality. Subsequently a correlation between the histology results and the imaging features (Radiologic-Histologic correlation) is performed to ensure adequate sampling of these lesions and avoid cancers being missed.[5] The majority of breast biopsies will yield definitive results.[6] However, in 5% to 15% of cases, the results are non-definitive,[5] and surgical excisional biopsy is recommended to determine the final pathology and rule out the presence of malignancy. If a malignancy is subsequently confirmed, the case is "upgraded" from non-definitive to malignant (approximately 10-20%). In the US, women over the age of 20 have an annual breast biopsy utilization rate of 62.6 per 10,000, translating to over 700,000 women undergoing breast core biopsy in 2010.[7] Knowing this, approximately 35,000 to 105,000 of these women likely underwent excision, an additional and more invasive procedure. Ultimately, a majority of these women receive a benign diagnosis.

Typical practice holds that all non-definitive cases should go on to excision for the benefit of the truly malignant cases. In the mid-1990s, the American College of Radiology developed the mammography lexicon, Breast Imaging Reporting and Data System (BI-RADS), to standardize mammogram feature distinctions and the terminology used to describe them,[8] and studies show that BI-RADS descriptors are predictive of malignancy,[9] specific histology,[10] and prognostic significance.[11] Given that 1) a complex combination of variables predict upgrade, 2) reliable data including accurate outcomes via cancer registries are available, and 3) accurate prediction of upgrade would substantially improve management makes this domain ripe for decision support that would have a substantial impact on patient care.

Perhaps the greatest challenge in this particular task is the relative rarity of the event. Though the number of women affected by non-definitive CNB across the US is substantial, the availability of data to any particular institution is relatively limited, thus making the development of decision support systems more difficult.

As such, the clinical objective, as determined by our medical experts, is to identify when a breast core biopsy may not accurately reflect existing breast disease and further investigation (with re-biopsy or surgical excision) will reveal malignancy, also called an "upgrade". We refer to this as the "upgrade" prediction task.

Application of ABLe to Upgrade Prediction

In our first meeting to establish a model, the MDEs defined the task (Step 1.0) as predicting upgrade using a dataset of 157 biopsies that were prospectively given a non-definitive diagnosis at radiologic-histologic correlation conference. To incorporate physician advice about relationships between variables (Step 2.0), we opted to use a logic-based language. Our variables consist of imaging findings, demographic information, and some pathology findings. Physicians and computer scientists hand-coded expert rules expressing combinations of variables that increase or decrease risk of upgrade according to physician experience and the literature. For example, the MDEs expressed that if the imaging abnormality presents a spiculated mass margin, or an irregular mass shape, or high mass density, or is increasing (e.g. increasing mass size, or number of calcifications), then the risk of upgrade increases. To incorporate the rules into the model, we took a similar approach to Dutra et al.,[12] treating them as binary features in our dataset. We chose to use the Naïve Bayes algorithm and valued false negatives equally to false positives (Step 2.0). We chose Naïve Bayes for its simplicity, making our modeling methods more approachable to non-experts, and for its history of good performance, despite strong independence assumptions[13].

The initial model (Step 3.0) demonstrated no substantial ability to identify benign non-definitive cases without misclassifying malignant cases. Our MDEs considered the results produced by this initial model (Step 1.1) and formed a hypothesis about what the primary challenge was. Specifically, they surmised that the non-definitive population as a whole was too challenging to be addressable with the predictive features available in a typical clinical dataset, and that targeting a subpopulation would be most fruitful.

Non-definitive biopsies can be broken into three subtypes (discordant, insufficient, and atypical/radial scar), and each subtype has distinct features that are likely to predict upgrade. Discordance means that the histologic findings do not provide an acceptable explanation for the imaging features and indicates that the targeted tissue may not have been sampled adequately. In this situation, the imaging features are likely to provide the most influential variables in predicting upgrade. The other two subtypes focus more on the histologic factors (cellular features) that raise the possibility that abnormal tissue remains in the region of the biopsy, in which case, pathology features should be more influential in predicting upgrade. The dataset available to us for the task contained a limited set of pathology features, but we use structured reporting in our imaging practice and adhere to the BI-RADS lexicon, so our experts identified an opportunity to employ machine learning methods that capitalize on the imaging features. Thus, we chose to alter the task (Step 2.1) and focus on estimating the probability of malignancy for discordant cases specifically. Due to the alteration of the task, our physician experts also reduced the initial set of variable relationship rules (Step 2.1) to a set of four variables specifically related to predicting discordance. This also led us to begin collecting a larger set of features in our medical practice for the sake of future work on the other subpopulations. We again tested the model with cross-validation (Step 3.1).

Evaluation of the next model (Step 1.2) demonstrated a marked improvement over the initial model (see Table 1). When trained with the combined base feature set and the binary advice rules, the model showed improvement over models trained on either feature set alone, though at the cost of missing malignancies. We thus considered value specifications that would help the model better address the clinical objective (Step 2.2). This led us to specify a skewed cost-ratio for FN:FP to suggest that the algorithm prefer to misclassify every benign case before misclassifying a single malignancy. We selected a skew of 50:1 based on the 50 benign cases in the discordant set (Step 2.2). We again tested the model with cross-validation (Step 3.2).

Review of the new model (Step 1.3) demonstrated an improvement over the previous model, but still misclassified a

Table 1: 10-fold cross-validated performance of Naïve Bayes classifiers with FN:FP cost-ratio of 1:1 and our final model with cost-ratio 150:1 at 2% threshold of excision.

		FN:FP cost-ratio 1:1			FN:FP cost-ratio 150:1		
Parameter	Baseline	Data	Rules	Data + Rules	Data	Rules	Data + Rules
Biopsy	60	28	42	30	55	55	48
No Biopsy	0	32	18	30	5	5	12
Malignant Excisions	10	7	9	7	10	10	10
Benign Excisions	50	21	33	23	45	45	38
PPV (%)	16.7	25.0	21.4	23.3	18.2	18.2	20.8
Specificity (%)	0.0	70.0	34.0	54.0	10.0	10.0	24.0
10-fold specificity > 0.0 p-Value	-	0.000$*$	0.003$*$	0.000$*$	0.026	0.026	0.004$*$

single malignant case. Upon inspection, this single misclassification was shown to be the result of incorrectly entered features in our reporting software. These kinds of errors should be expected in real-world data, which led us to reconsider our skewed cost-ratio (Step 2.3). Our MDEs suggested a more conservative cost-ratio of 150:1 based not just on counts in the dataset, but on recent work on utility analysis in mammography[14] (Step 3.3). This led us to our final model, which demonstrated our best performance (see Table 1).

Materials and Methods

Our example study included a population of patients that underwent 1,910 consecutive CNB, as a result of a diagnostic mammogram, from Jan 1, 2006 to Dec 31, 2011. A total of 157 biopsies were prospectively given a non-definitive diagnosis at radiologic-histologic correlation conference, and 60 of these were categorized as discordant. Recall that we have chosen to focus on the discordant cases. The mean age of these patients was 55.2 years (range= $25 - 83$ years, sd= 12.2), all 60 cases were women, and all underwent excision. As a reference standard for final diagnosis, we use the result of excisional biopsy (within 6 months after CNB) or a registry match (within 1 year after CNB). 50 were confirmed to be benign while 10 (16.7%) were found to be malignant.

All mammographic findings were described and recorded using BI-RADS terms by the interpreting radiologist at the time of mammography interpretation using structured reporting software (PenRad®, Minnetonka, MN), which is routinely used in the clinical practice. We derive mammography features and demographic risk factors from the diagnostic mammogram that precedes the biopsy and has an abnormal BI-RADS assessment category. All expert advice rules are translated into binary features that are either included in or excluded from the dataset at training time.

We use 10-fold stratified cross-validation for evaluation, and the Weka[15] software package (version 3.7) to train a Naïve Bayes (NB) model for each fold. Note that NB is known to often accurately predict the most probable label even though its predicted probabilities are not well calibrated.[16] Based on our clinical objective we specify that the cost of misclassifying a malignant case is greater than that of misclassifying a benign case. We show results with a FN:FP cost-ratio of 1:1 and a suggested cost ratio of 150:1 drawn from the literature.[14]

For final evaluation of our trained models, we merge the output probabilities on the test portions of all folds. We compare our trained models to the baseline of current standard practice. For comparison, we consider a 2% threshold of predicted malignancy that would hypothetically be used to decide if the patient should go on to excisional biopsy. This threshold has been used previously and is clinically reasonable.[17] The baseline current practice is to excise all discordant cases, so there is no distinction of treatment at any threshold. For each model, we compare the number of malignant cases that would be excised, the number of benign cases that would be excised, the positive predictive value (PPV) of malignancy, and specificity. We use a one-sided, one-sample t-test at the 99% confidence level to compare the specificity of the NB classifier to the baseline specificity of 0.0 (i.e. excision of all cases).

Results

The results produced by our final model are in the rightmost block of Table 1. This shows the performance of our NB classifier trained on our discordant cases, with a FN:FP cost-ratio of 150:1, and a threshold for excision with model output greater than or equal 0.02. The first column of results shows the baseline performance for comparison, while the subsequent blocks show results when trained on just the data, just the binary advice features, and the two combined, at different cost-ratios. The left block of Table 1 shows results when the cost-ratio is set to 1:1. We do not present results of training on all non-definitive biopsies, but initial experiments showed no significant ability to reduce benign excisions at an output threshold of 0.02.

Conclusion

We present a framework called ABLe for incorporating expert clinical knowledge into machine learning models for decision support. The framework consists of three different categories of advice that are used to iteratively refine model development. We describe ABLe in detail and illustrate its application to an example task. In this example task, we train Naïve Bayes models to estimate the probability of upgrade following a discordant core needle biopsy. We train our models using our base dataset, using just advice-based features, and the two combined. Additionally, we train our models with costs skewed such that misclassifying benign cases is far preferable to misclassifying even a single malignancy. We achieved our best results by applying ABLe to the task. Furthermore, our results suggest that we can significantly reduce the number of truly benign discordant cases that go on to excision without missing a single malignancy. We find these results and the incorporation of expert knowledge very promising, and are actively seeking more data to improve performance and to further validate our methods.

Acknowledgements

The authors acknowledge support from the National Institutes of Health (grants: R01LM011028 and R01CA165229) and from FCT-Portugal (grant PTDC/EEI-SII/2094/2012).

References

[1] Lucas P. Expert Knowledge and Its Role in Learning Bayesian Networks in Medicine: An Appraisal. In: Quaglini S, Barahona P, Andreassen S, editors. Artificial Intelligence in Medicine. vol. 2101 of Lecture Notes in Computer Science. Springer Berlin Heidelberg; 2001. p. 156–166.

[2] Mitchell TM. Machine Learning. New York: McGraw-Hill; 1997.

[3] Gibert K, García-Alonso C, Salvador-Carulla L. Integrating clinicians, knowledge and data: expert-based cooperative analysis in healthcare decision support. Health Res Policy Syst. 2010;8(28):1478–4505.

[4] Velikova M, Lucas PJF, Samulski M, Karssemeijer N. On the interplay of machine learning and background knowledge in image interpretation by Bayesian networks. Artificial Intelligence in Medicine. 2013;57(1):73 – 86. Available from: `http://www.sciencedirect.com/science/article/pii/S0933365712001522`.

[5] Liberman L. Percutaneous imaging-guided core breast biopsy: state of the art at the millennium. American Journal of Roentgenology. 2000;174(5):1191–1199.

[6] Bruening W, Fontanarosa J, Tipton K, Treadwell JR, Launders J, Schoelles K. Systematic review: comparative effectiveness of core-needle and open surgical biopsy to diagnose breast lesions. Annals of Internal Medicine. 2010 Feb, 16th;152(4):238–246. Epub 2009 Dec 14.

[7] National Vital Statistics Report. National Center for Health Statistics; 1998.

[8] Breast Imaging Reporting and Data System (BI-RADS^{TM}). Reston, VA, USA; 2003.

[9] Liberman L, Abramson A, Squires F, Glassman J, Morris E, Dershaw D. The Breast Imaging Reporting and Data System: positive predictive value of mammographic features and final assessment categories. American Journal of Roentgenology. 1998;171:35–40.

[10] Burnside E, Rubin D, Shachter R, Sohlich R, Sickles E. A probabilistic expert system that provides automated mammographic-histologic correlation: initial experience. American Journal of Roentgenology. 2004;182:481–488.

[11] Tabar L, Chen T, Yen A, Tot T, Tung T, Chen L, et al. Mammographic tumor features can predict long-term outcomes reliably in women with 1-14-mm invasive breast carcinoma. Cancer. 2004;101(8):1745–1759.

[12] Dutra I, Nassif H, Page D, Shavlik J, Strigel R, Wu Y, et al. Integrating machine learning and physician knowledge to improve the accuracy of breast biopsy. In: AMIA Annual Symposium Proceedings. Washington, DC; 2011. p. 349–355.

[13] Zhang H. The optimality of naive Bayes. In: Proceedings of the FLAIRS Conference; 2004. p. 3–9.

[14] Abbey C, Eckstein M, Boone J. Estimating the relative utility of screening mammography. Medical Decision Making. 2013;33(4):510–520.

[15] Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten I. The WEKA Data Mining Software: an update. SIGKDD Explorations Newsletter. 2009;11(1):10–18.

[16] Zadrozny B, Elkan C. Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers. In: In Proceedings of the Eighteenth International Conference on Machine Learning. Morgan Kaufmann; 2001. p. 609–616.

[17] Burnside E, Davis J, Chhatwal J, Alagoz O, Lindstrom M, Geller B, et al. Probabilistic computer model developed from clinical data in national mammography database format to classify mammographic findings. Radiology. 2009;251:663–672.