Score As You Lift (SAYL): A Statistical Relational Learning Approach to Uplift Modeling

Houssam Nassif[1], Finn Kuusisto[1], Elizabeth S. Burnside[1], David Page[1], Jude Shavlik[1], and Vítor Santos Costa[1,2]

[1]University of Wisconsin, Madison, USA

[2]CRACS-INESC TEC and DCC-FCUP, University of Porto, Portugal

Abstract

We introduce Score As You Lift (SAYL), a novel Statistical Relational Learning (SRL) algorithm, and apply it to an important task in the diagnosis of breast cancer. SAYL combines SRL with the marketing concept of uplift modeling, uses the area under the uplift curve to direct clause construction and final theory evaluation, integrates rule learning and probability assignment, and conditions the addition of each new theory rule to existing ones.

Breast cancer, the most common type of cancer among women, is categorized into two subtypes: an earlier in situ stage where cancer cells are still confined, and a subsequent invasive stage. Currently older women with in situ cancer are treated to prevent cancer progression, regardless of the fact that treatment may generate undesirable side-effects, and the woman may die of other causes. Younger women tend to have more aggressive cancers, while older women tend to have more indolent tumors. Therefore older women whose in situ tumors show significant dissimilarity with in situ cancer in younger women are less likely to progress, and can thus be considered for watchful waiting.

Motivated by this important problem, this work makes two main contributions. First, we present the first multi-relational uplift modeling system, and introduce, implement and evaluate a novel method to guide search in an SRL framework. Second, we compare our algorithm to previous approaches, and demonstrate that the system can indeed obtain differential rules of interest to an expert on real data, while significantly improving the data uplift.

SAYL

SAYL is a Statistical Relational Learner based on SAYU that integrates uplift modeling with the search for relational rules. Similar to SAYU, every valid rule generated is used to construct a Bayesian network (alongside with current theory rules) via propositionalization, but instead of constructing a single classifier, SAYL constructs two TAN classifiers; one Bayes net for each of the subject and control groups. Both classifiers use the same set of attributes, but are trained only on examples from their respective groups. SAYL uses the TAN generated probabilities to construct the lift and uplift curves, where area under the uplift curve (AUU) is the difference in areas under the lift curves (AUL).

Algorithm SAYL

```
Rs ← {}; M_0^s, M_0^c ← InitClassifiers(Rs)
while DoSearch() do
    e_s^+ ← RandomSeed();
    ⊥_{e_s^+} ← saturate(e);
    while c ← reduce(⊥_{e_s^+}) do
        M^s, M^c ← LearnClassifiers(Rs ∪ {c});
        if Better(M^s, M^c, M_0^s, M_0^c) then
            Rs ← Rs ∪ {c}; M_0^s, M_0^c ← M^s, M^c;
            break
        end if
    end while
end while
```

Acknowledgements

This work is supported by US National Institute of Health (NIH) grant R01-CA165229. VSC was funded by the ERDF through the Progr. COMPETE, the Portuguese Gov. through FCT, proj. ABLe ref. PTDC/EEI-SII/2094/2012, ADE (PTDC/ EIA-EIA/121686/2010), and by FEDER/ON2 and FCT project NORTE-07-124-FEDER-000059.

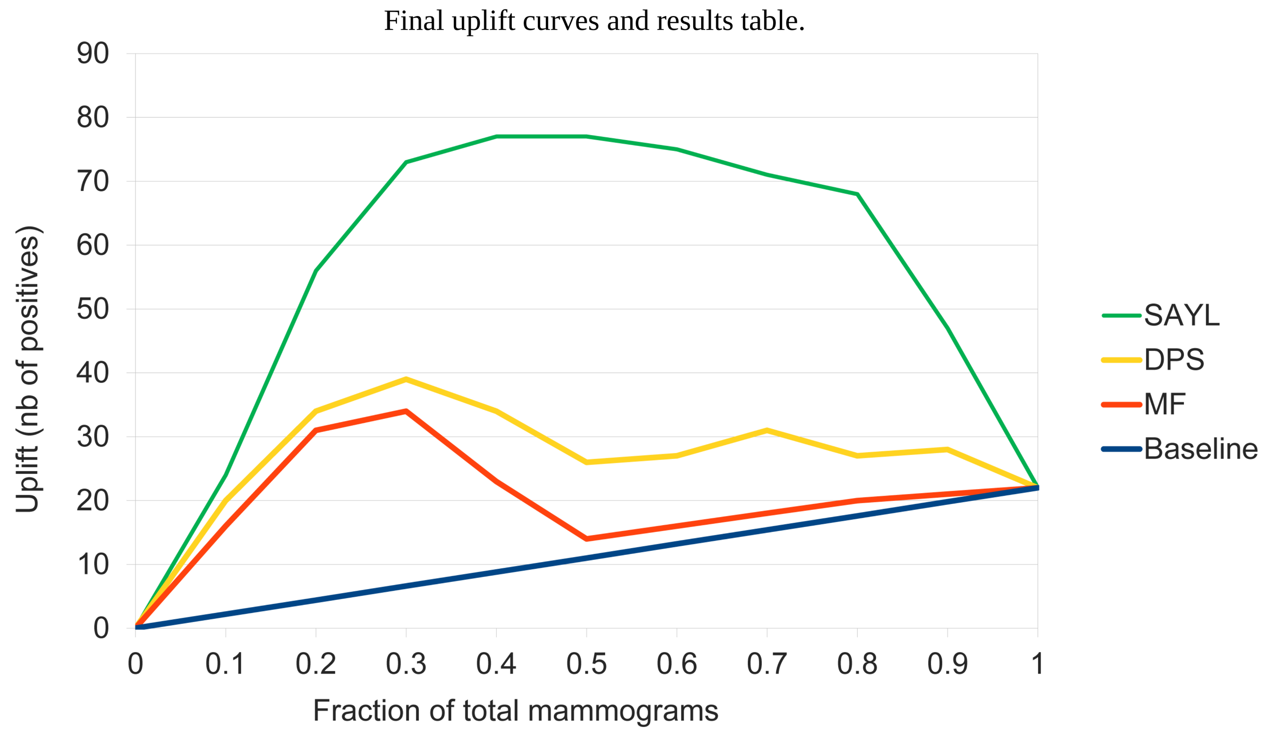
Results

We use 10-fold cross-validation, making sure all records pertaining to the same patient are in the same fold. For each cross-validated run, we use 4 training, 5 tuning and 1 testing folds. For each fold, we used the best combination of parameters according to a 9-fold internal cross-validation using 4 training, 4 tuning and 1 testing folds. We concatenate the results of each testing set to generate the final uplift curve.

Composition of our dataset.

Older		Younger	
In Situ	Invasive	In Situ	Invasive
132	401	110	264

Final uplift curves and results table.



Algorithm	Uplift AUC	Lift (Older) AUC	Lift (Younger) AUC	Rules Avg #	DPS p-value	
SAYL	58.10	97.24	39.15	9.3	0.0020	*
DPS	27.83	101.01	73.17	37.1	-	
MF	20.90	100.89	80.99	19.9	0.0039	*
Baseline	11.00	66.00	55.00	-	0.0020	*

We use the Mann-Whitney test at 95% confidence to compare experiments. SAYL produces much greater uplift and significantly outperforms previous ILP-based methods (indicated by *), as well as the baseline of random guessing.

Model Interpretation

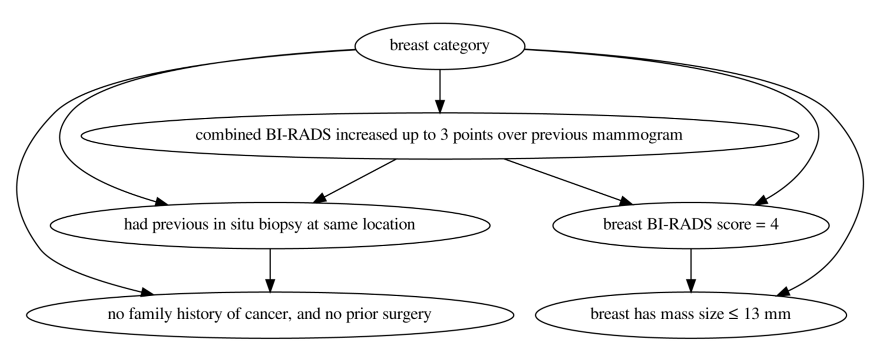
SAYL returns two TAN Bayes-net models, one for the older and one for the younger, with the same first-order logic rules as the nodes. Although both models have the same rules as nodes, TAN learns the structure of each model on its corresponding data subset separately, resulting in different networks. SAYL identifies the features that best differentiate amongst subject and control positive examples, while TAN uses these features to create the best classifier over each set.

For example, SAYL produced the following rules on one run:

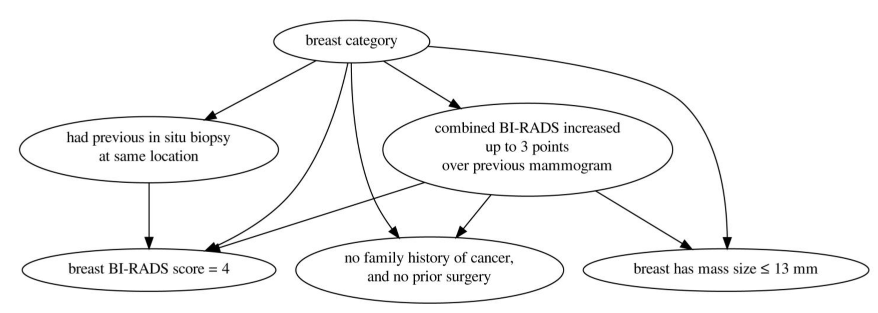
1. Current study combined BI-RADS increased up to 3 points over previous mammogram.
2. Patient had previous in situ biopsy at same location.
3. Breast BI-RADS score = 4.

Not only are these rules chosen by SAYL to maximize uplift, but they capture themes that have been identified in previous work as differentiating older and younger patients.

Example TAN model for older subgroup.



Example TAN model for younger subgroup.



Uplift Modeling

Uplift modeling is a predictive modeling technique that attempts to specifically characterize a target subgroup of a population over a control subgroup. This is accomplished by maximizing the uplift of a model, or models, over the two subgroups

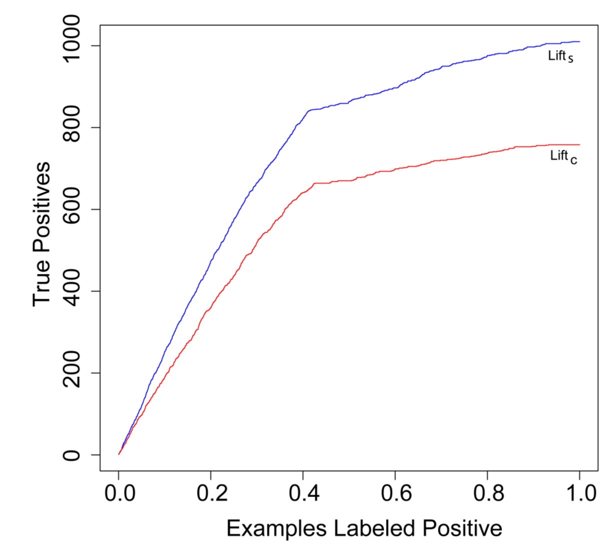
$$Uplift(M_s, M_c, \rho) = Lift_{M_s}(S, \rho) - Lift_{M_c}(C, \rho)$$

$$AUL = \int Lift(D, \rho)\, d\rho$$

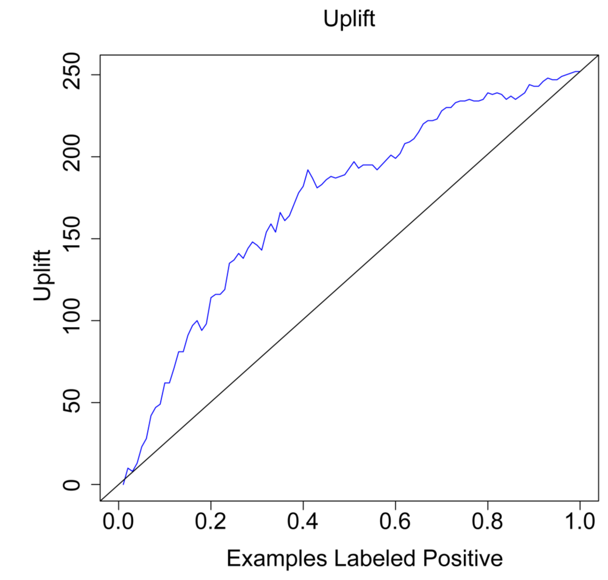
$$\approx \frac{1}{2} \sum_{k=1}^{P+N} (\rho_{k+1} - \rho_k)\big(Lift(D, \rho_{k+1}) + Lift(D, \rho_k)\big)$$

$$AUU = AUL_{M_s} - AUL_{M_c}$$

Lift



Uplift



Conclusions

We present Score As You Lift (SAYL), a novel Statistical Relational Learning algorithm and the first multi-relational uplift modeling system. Our algorithm maximizes the area under the uplift curve, uses this measure during clause construction and final theory evaluation, integrates rule learning and probability assignment, and conditions the addition of new theory rules to existing ones. SAYL significantly outperforms previous approaches on a mammography application (p = 0.002 with similar parameters), while still producing human interpretable models. We plan on further investigating the clinical relevance of our model, and to apply SAYL to additional differential problems.

Selected References

Davis, J., Burnside, E., de Castro Dutra, I., Page, D., Santos Costa, V.: An integrated approach to learning Bayesian Networks of rules. In: Proceedings of the 16th European Conference on Machine Learning. pp. 84–95. Porto, Portugal (2005).

Hansotia, B., Rukstales, B.: Incremental value modeling. Journal of Interactive Marketing 16(3), 35–46 (2002).

Nassif, H., Kuusisto, F., Burnside, E., Page, D., Shavlik, J., Santos Costa, V.: Score as you lift (SAYL): A statistical relational learning approach to uplift modeling. In: European Conference on Machine Learning (ECML-PKDD). Prague (2013).

Nassif, H., Santos Costa, V., Burnside, E., Page, D.: Relational differential prediction. In: European Conference on Machine Learning (ECML-PKDD). pp. 617–632. Bristol, UK (2012).

Radcliffe, N. and Simpson. R.: Identifying who can be saved and who will be driven away by retention activity. Journal of Telecommunications Management 1(2), 168–176 (2008).

Radcliffe, N., Surry, P.: Differential response analysis: Modeling true response by isolating the effect of a single action. In: Credit Scoring and Credit Control VI. Edinburgh, Scotland (1999).

Radcliffe, N., Surry, P.: Real-world uplift modelling with significance-based uplift trees. White Paper TR-2011-1, Stochastic Solutions (2011).

Rzepakowski, P., Jaroszewicz, S.: Decision trees for uplift modeling with single and multiple treatments. Knowledge and Information Systems 32, 303–327 (2012).

Tuffery, S.: Data Mining and Statistics for Decision Making. John Wiley & Sons, 2nd edn. (2011).