Relational Differential Prediction

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

[1] University of Wisconsin, Madison, USA
[2] University of Porto, Portugal

Abstract. A typical classification problem involves building a model to correctly segregate instances of two or more classes. Such a model exhibits differential prediction with respect to given data subsets when its performance is significantly different over these subsets. Driven by a mammography application, we aim at learning rules that predict breast cancer stage while maximizing differential prediction over age-stratified data. In this work, we present the first multi-relational differential prediction (aka uplift modeling) system, and propose three different approaches to learn differential predictive rules within the Inductive Logic Programming framework. We first test and validate our methods on synthetic data, then apply them on a mammography dataset for breast cancer stage differential prediction rule discovery. We mine a novel rule linking calcification to *in situ* breast cancer in older women.

Keywords: Uplift modeling, relational data mining, differential prediction, inductive logic programming, ILP, stratified data, breast cancer, in situ.

1 Introduction

A recurrent problem in social sciences is to understand why two or more different populations exhibit differences in a trait. In psychology [8,20,36], one may want to assess the fairness of a test over several different populations. In marketing [17,27,21], one may want to compare subjects and controls in order to study the effectiveness of an advertising campaign. Similar tasks thus arise in several domains and depending on the domain, the problem is known as *differential prediction*, *differential response analysis*, or *uplift modeling*.

In contrast to most studies of *differential prediction* in psychology, marketing's *uplift modeling* assumes an active agent. But, given that in both cases we have two populations that have been subjected to an external agent, we argue that the concepts and techniques originally developed for uplift marketing can, and should, apply to the task of differential prediction (and vice versa). Differential prediction has been studied extensively in the context of multi-attribute data [30,28]. One approach is to generate different classifiers for each sub-population, and to look for differences between the classifiers. Further progress requires building models driven by evaluation functions that take into

P. Flach et al. (Eds.): ECML PKDD 2012, Part I, LNCS 7523, pp. 617–632, 2012.

account the differential nature of uplift modeling [29]. Also, techniques such as *uplift curves* have made it possible to evaluate and compare differential models.

An important differential problem arises in the area of breast cancer research. Breast cancer is the most common type of cancer among women, with a 12% probability of incidence in a lifetime [3]. Breast cancer has two basic stages: an earlier *in situ* stage where cancer cells are still confined where they developed, and a subsequent *invasive* stage where cancer cells infiltrate surrounding tissue. Since nearly all in situ cases can be cured [2], current practice is to treat in situ occurrences in order to avoid progression into invasive tumors [3]. Nevertheless, the time required for an in situ tumor to reach invasive stage may be sufficiently long for a woman to die of other causes; raising the possibility that the diagnosis and treatment may not have been necessary, a phenomenon called *overdiagnosis*.

Cancer occurrence and stage are determined through biopsy, a costly, invasive, and potentially painful procedure. Actual treatment is costly, and may generate undesirable side-effects. For these reasons, the 2009 US National Institutes of Health consensus conference on ductal carcinoma in situ highlighted the need for methods that can accurately identify patient subgroups that would benefit most from treatment, as well as those who do not need treatment [1]. In recent work, Nassif *et al.* [25] reported that different pre-biopsy mammographic features can indeed be used to classify cancer as invasive or in situ *for different age groups*. They identified invasive/in situ classification rules that have significantly different performance across age strata. This finding confirms that, based on age, different mammographic features can be used to classify cancer stage. The key motivation to this work is *to understand how breast cancer evolves differently across different age groups, and what features exhibit differential cancer stage prediction across age*.

Differential breast cancer prediction introduces two novel problems to differential prediction. First, in order to classify a sample, best results require taking into account previous and simultaneous samples for the same patient [9]. This demands a multi-relational data representation. We thus need a *relational* differential model. Second, it is of utmost importance that experts be able to interpret the results and identify patient subgroups. Both challenges can be addressed by using *rules* to represent the model.

We hereby introduce a rule-based multi-relational differential classifier, and demonstrate its applicability on medical data. This work makes three main contributions. First, we present the first multi-relational differential modeling system, and introduce, implement and evaluate novel methods to guide search in a rule-based differential setting. We propose three general methods that are implemented within the Inductive Logic Programming (ILP) framework [23,11], a commonly used approach for relational data mining. We opt for ILP-based rule learning instead of decision-tree-based rule learning because the latter is a special case of the former [6,34]. Second, we present a detailed evaluation of the applicability and usefulness of our approach under different data sizes and noise rates through simulated data. Third, we demonstrate that the system can indeed obtain differential rules of interest to an expert on real data.

2 Related Work

To the best of our knowledge, differential prediction was first used in psychology to assess the fairness of cognitive and educational tests. In this area, it is defined as the case where consistent nonzero errors of prediction are made for members of a given subgroup [8], and it is detected by fitting a common regression equation and checking for systematic prediction discrepancies for given subgroups, or by building regression models for each subgroup and testing for differences between the resulting models [20,36]. The standard approach uses moderated multiple regression, where the criterion measure is regressed on the predictor score, subgroup membership, and an interaction term between the two [5,33]. If the predictive model differs in terms of slopes or intercepts, it implies that bias exists because systematic errors of prediction would be made on the basis of group membership.

An example is assessing how college admission test scores predict first year cumulative grades for males and females. For each gender group, we fit a regression model. We then compare the slope, intercept and/or standard errors for both models. If they differ, then the test exhibits differential prediction and may be considered unfair.

The same concept arises in case-control studies, and is referred to as *differential misclassification.* Instances are cross-classified by case-control status and exposure category. An exposure misclassification is defined as differential if the probabilities of misclassification differ for instances with different case-control categories. Similarly, a case-control misclassification is defined as differential if the probabilities of misclassification differ for instances with different exposure categories [7,13]. This concept is the basis of the related machine learning concept of "differential misclassification cost", incorporating different misclassification costs into a cost sensitive classifier [31].

An important application of differential prediction is in marketing studies, where it can be used to understand the best targets for an advertising campaign and it is often known as uplift modeling. Seminal work includes Radcliffe and Surry's true response modeling [27], Lo's true lift model [21], and Hansotia and Rukstales' incremental value modeling [17]. As an example, Hansotia and Rukstales construct a regression and a decision tree, or CHART, model to identify customers for whom direct marketing has sufficiently large impact. The splitting criterion is obtained by computing the difference between the estimated probability increase for the attribute on the treatment set and the estimated probability increase on the control set.

Recent work by Rzepakowski and Jaroszewicz [29] suggests that performance of a tree-based uplift model may improve by using a divergence statistic. The authors propose three postulates that should be obeyed by tree-based splitting criteria. First, the value of the splitting criterion is minimum if and only if the class distributions in treatment and control groups are the same in all branches. Second, splitting criterion is zero if treatment and control are independent. Third, if the control group is empty, the criterion reduces to the case measure. They introduce two new statistics, one based on Kullback-Leibler

divergence, the other based on Euclidean distance. Evaluation on prepared data suggests improved performance. Radcliffe and Surry [28] criticize one of the postulates and the fact that the measures are independent of population size, a parameter that they consider crucial in practical applications.

We observe that the task of discriminating between two dataset strata is closely related to the problem of Relational Subgroup Discovery (RSD), that is, "given a population of individuals with some properties, find subgroups that are statistically interesting" [37]. In the context of multi-relational learning systems, RSD applies a first propositionalization step and then applies a weighted covering algorithm to search for rules that can be considered to define a sub-group in the data. Although the weighting function is defined to focus on unexplored data by decreasing the weight of covered examples, RSD does not explicitly aim at discovering the differences between given partitions.

3 Differential Predictive Concept Definition

Given data that can be partitioned into a set of strata, we define a differential predictive concept as a concept whose measure is significantly different over one stratum as compared to the others. To be more precise, we define a stratified dataset as one composed of disjoint partitions, where each partition contains at least one instance of each target class.

Definition 1 (Stratified Dataset). *Let tc be a target class defined over the set of instances X, and let $D = \{\langle x, tc(x)\rangle\}$ be a set of training examples labeled according to tc. Let $\{D_1, \ldots, D_n\}$ be n disjoint subsets of D, and let D_i^l be the set of training examples of D_i with class label l, such that:*

$$(\forall (i,j) \in [1,n],\ i \neq j)\ D_i \subset D,\ D_i \cap D_j = \emptyset,\ \forall l\ D_i^l \neq \emptyset. \tag{1}$$

A k-strata dataset $\mathcal{D}$ over the set of instances X is the union of k such subsets D_i, with $2 \leq k \leq n$, such that:

$$\mathcal{D} = \{D_i \mid 1 \leq i \leq k\}. \tag{2}$$

After specifying the instance space, we define a differential predictive concept.

Definition 2 (Differential Predictive Concept). *Let c be a concept over the set of instances X, and let $\mathcal{D}$ be a k-strata dataset. Let $S(c|D_i)$ be the classification performance score for c over the subset D_i. A* stratum-j specific differential predictive concept *is a concept c_j such that:*

$$\forall i \neq j,\ S(c_j|D_j) \gg S(c_j|D_i). \tag{3}$$

Score difference ($\gg$) can be evaluated using statistical significance tests or by comparing against a threshold. In this work we will focus on 2-strata 2-class differential problems.

4 Learning Differential Predictive Rules

This work uses Inductive Logic Programming (ILP) [11] to build the first relational differential classifier. The benefit of using ILP in this context is twofold. First, we can use a first-order logic formulation to represent complex relational patterns spanning the patient and mammogram levels. In our motivating application, we can represent data on one mammogram and relate it to prior mammograms for the same patient. Second, we shall take advantage of ILP's ability to learn easily-comprehensible logical rules.

Used for differential prediction, ILP — as a rule-learning technique — has a major advantage: each individual rule can be viewed as a feature describing a subgroup. We can investigate the performance of each rule on a given dataset, identify rules that only apply to particular data subsets, and isolate subgroups covered by a particular rule. Given a stratified dataset, we can examine the performance of rules on the various strata, and select stratum-specific rules that have significantly different performances across strata.

We propose and evaluate three different approaches to learn differential predictive rules. All three approaches can be applied to any ILP algorithm, and can be used with any scoring function S. We use m-estimate to represent the probability of an example given a rule. We set both m and the minimum number of positive examples to be covered by an acceptable clause to 10% of the number of positive examples per stratum and class.

An important concern in real-life situations is population size [28]. Probability estimates tend to favor highly precise estimates (even taking into account the m count) and may be prone to overfitting, a difficult problem in ILP given the number of rules we generate and their complexity. In this work, we heuristically compensate for population size by weighing over the rule positive cover on the case set, as shown below.

4.1 Baseline Approach

As a running example, suppose we are given a 2-strata 2-class dataset of breast cancer records, with class labels *in_situ* and *invasive*, and strata *older* and *younger*. Our task is to find rules that exhibit a differential performance over the two strata. More precisely, we want rules that correctly predict *in_situ* versus *invasive* in the older stratum, but have a significantly worse performance over the younger stratum. Our target stratum D_t is thus *older*, while *younger* is the other stratum D_o.

A simple approach is to merge both strata together while including the stratifying attribute as an additional predicate in the background knowledge. Thus older stratum examples will have $stratum(Example, \texttt{older})$ as an additional feature, while $stratum(Example, \texttt{younger})$ will describe younger instances. We run ILP over the whole dataset and select theory rules that have the condition $stratum(Example, \texttt{older})$ in their body. Such rules are specific to the older stratum. We call this approach the *baseline* approach (BASE).

We score each rule R by considering its positive cover and m-estimate over the merged strata:

$$S_{BASE}(R|D_t, D_o) = poscover(R|D_t \cup D_o) \times mestimate(R|D_t \cup D_o). \quad (4)$$

4.2 Model Filtering Approach

Our second method is a *model filtering* (MF) approach based on [25]. It follows similar principles to the Two Model approach [21,28]. We start by constructing a predictive ILP model over a given stratum. The model outputs a high-performance stratum-specific theory. By construction, the theory rules perform well on their stratum, according to a given scoring function S. We test each theory rule on the other stratum, and select rules with a poor performance, hence filtering the original model. According to this model, the greater the performance difference, the more differential predictive a rule should be.

Fig. 1 flowchart outlines the construction of in situ rules specific to the `older` stratum. Starting with the older subset, we construct an ILP model that discriminates between *in_situ* and *invasive*. The generated rules are expected to have good performance over the `older` stratum. We then test each rule on the `younger` stratum, and keep rules that perform poorly.

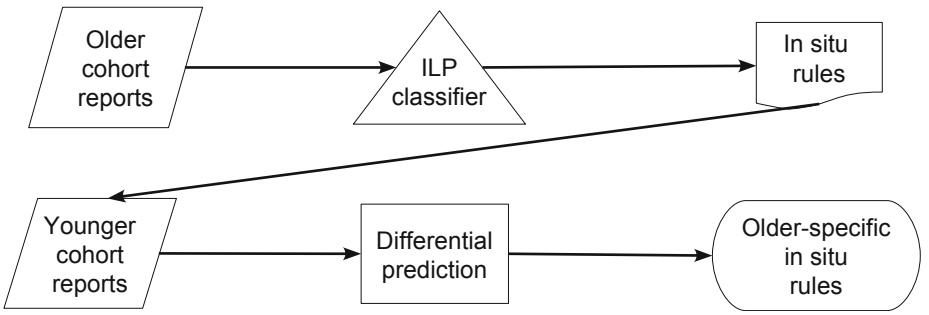


Fig. 1. Model Filtering approach to identify older-specific in situ rules

During the MF search phase, we score a rule R over strata D_t using $S_{BASE}(R|D_t)$. Given the final theory, we score each theory rule R_t according to:

$$S_{MF}(R_t|D_t, D_o) = S_{BASE}(R_t|D_t) - S_{BASE}(R_t|D_o). \quad (5)$$

4.3 Differential Prediction Search Approach

Our third method, *differential prediction search* (DPS), builds a differential prediction ILP classifier by altering the ILP search. Unlike our generate-then-test model filtering method, DPS uses *test-incorporation* by altering the ILP search space. It defines a new clause evaluation function that considers both strata

during search-space exploration and rule construction. This allows ILP to return rules specifically selected for their differential prediction score, that it would have overlooked otherwise. This is achieved through a differential-prediction-sensitive score that measures the performance difference of a rule over both strata.

Definition 3 (Differential-Prediction-Sensitive Scoring). *Let R be a clause (rule) over the set of instances X, and let $\mathcal{D}$ be a 2-strata dataset over X. We define a* differential-prediction-sensitive scoring function *Q as a function of R, D_t and D_o, such that Q is positively correlated to the performance of R over D_t, and negatively correlated to the performance of R over D_o.*

For the DPS method, we introduce the following differential-prediction-sensitive scoring function:

$$Q_{DPS}(R|D_t, D_o) = poscover(R|D_t) \times (mestimate(R|D_t) - mestimate(R|D_o)). \tag{6}$$

Note that this function is non-monotonic, as are most user-defined scoring functions, which prohibits us from custom-pruning the search space.

It is enlightening to relate this scoring function with the postulates described in [29]. Postulate 2 is trivially satisfied: if the condition is independent from treatment than the measure should indeed be zero. In contrast to postulate 1, we select rules that do *better* in one strata, and not rules that do *differently*. This is standard in ILP, where the search aims at covering the positive examples, E^+. In fact, in this setting, the standard techniques to explain negatives is to perform another search, switching E^+ and E^-. The last postulate concerns the case where the control set is empty. In this case, this measure indeed reduces to a classic non-differential ILP scoring function.

Our work thus obeys the main postulates followed by prior work in uplift modeling. Regardless, we observe that, to the best of our knowledge, this the first approach directly designed to learn differential rules. Instead, prior work on differential prediction has focused on learning trees or logistic regression models that can estimate differential performance. Instead, our work focuses on understanding factors that describe differential performance.

Fig. 2 flowchart outlines the construction of older-specific in situ rules. The differential-prediction classifier takes both strata as input. It constructs, scores and selects rules according to their differential-prediction-sensitive score.

5 Experimental Setting

We implement our three differential predictive rule learning methods using Aleph [34]. We invoke *induce_max*, which induces a theory that is unaffected by the order of the examples. We set $depth = 100000$, $i = 10$, $nodes = 50000$ and $clauselength = 5$. We perform experiments with the YAP Prolog compiler [32].

When using synthetic data, we know the ground truth. We then can compare the predicted rules to the original rules. We consider identical rules (up to variable renaming) as true findings. We label the remaining theory rules as

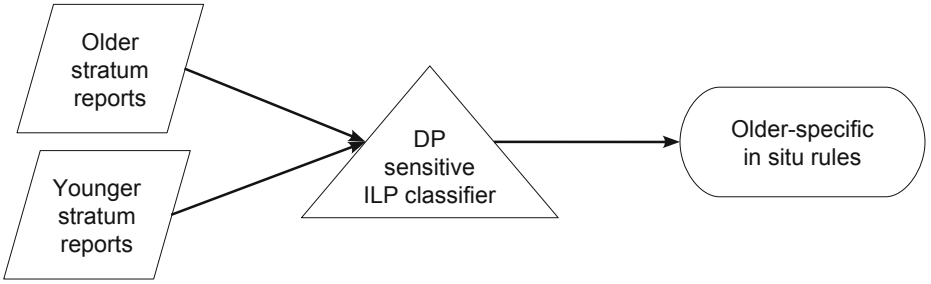


Fig. 2. Differential prediction search approach to identify older-specific in situ rules

false positive findings, and the missing original rules as false negative findings. We rank the theory rules by their score, and compute their precision-recall (PR) curve using [10]. Since we do not have scores associated with the missing false negative findings, we truncate the PR curve at the recall returned by the theory. Note that this yields a PR curve on recovered rules rather than on data.

We compare the different classifiers using their PR area under the curve (AUC-PR). We use the Mann-Whitney test to compare two sets of experiments. When comparing multiple sets, we use the Friedman test with a Hommel adjusted two-tailed Wilcoxon for the post-hoc pairwise tests. We chose these tests based on the recommendation of [12]. We set the confidence level to 95%.

Lacking differential rule ground truth, we can not use this method for real world data. Uplift curves are often used to address this problem [29]. Using 5-folds cross-validation, we use the learned theory rules as attributes to a TAN classifier [14] to assign a probability to each example. Given a threshold p, we compute the lift L_i, defined as the number of positive examples amongst the fraction p of examples that are ranked the highest on strata i. We generate an uplift curve by ranging p from 0 to 1 and plotting $\{p, L_1 - L_2\}$.

6 Synthetic Dataset

Before going to our target application, we use synthetic data to evaluate the ability of our approaches to uncover ground truth differential rules, and to study their sensitivity to variations in noise and in dataset size, two major concerns in real-world data. The multi-relational Michalski-trains dataset [19] is often used by ILP researchers to evaluate system performance in a controlled environment. Given two sets of trains, eastbound and westbound, the original problem consists of finding a concept which explains the eastbound trains. Each train includes multiple carriages of varying size, content and shape. Concept complexity is parametrized by generating more complex explanations of eastbound trains.

To test for differential prediction, we define two categories of trains, *red* and *blue*. We thus have a 2-strata (*red*, *blue*) 2-class (*east*, *west*) dataset. We randomly create up to 5 eastbound rules that are common for both *red* and *blue* trains. We then randomly create two additional sets of eastbound rules, each set

is specific to one stratum, *red* or *blue*. These are color-specific eastbound differential predictive rules. We ensure that all rules are unique, and that color-specific rules are not subsets of common rules nor of each other.

We generate the eastbound trains using the stratum's common and specific rules. We define westbound trains as non-eastbound trains. Our aim is to recover the color *red* differential predictive eastbound rules. They are our target rules.

As an example, suppose we have the following eastbound rules. Common eastbound rule:

$$east(T) \text{ :- } infront(T, C1, C2), short(C1), long(C2). \tag{7}$$

Stratum *red* specific eastbound rule (target rule):

$$east(T) \text{ :- } has\_car(T, C), jagged(C). \tag{8}$$

Stratum *blue* specific eastbound rule:

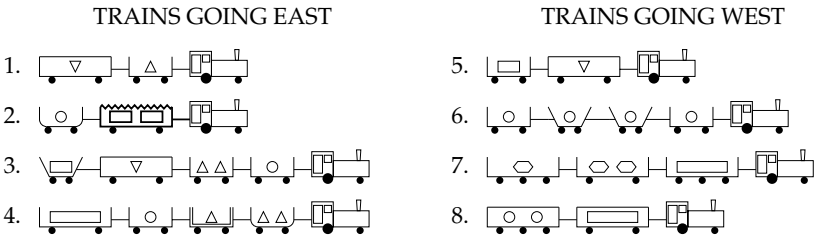
$$east(T) \text{ :- } has\_car(T, C), double(C). \tag{9}$$

Fig. 3(a) shows *red* trains, where eastbound trains 1, 3 and 4 have a short carriage in front of a long one (common rule), while train 2 has a jagged roof carriage (*red* specific rule). Fig. 3(b) shows *blue* trains, where eastbound trains 3 and 4 follow the common rule, while trains 1 and 2 have a double-hulled carriage (*blue* specific rule). Note a jagged roof on *blue* westbound train 5, it would have been classified eastbound if it was *red*.

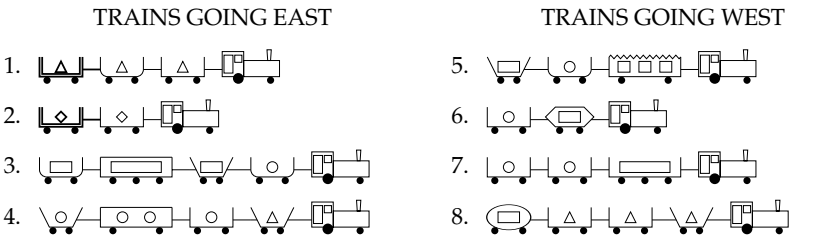
We devise two scenarios, the first with one *red* target rule to recover, and the second with up to 5 *red* target rules. For both scenarios we have up to 5 *blue*-specific rules. For each scenario, we randomly generate 30 different 2-strata 2-class train problems. For every problem, we use a random train generator [24] to randomly construct 1000 eastbound and 1000 westbound trains for each strata, for a total of 4000 trains per experiment. We ensure that each *red* eastbound target rule covers at least 10% of the eastbound *red* trains. We refer to this noise-free data as *clean1000*. To test the scalability of our algorithms, we also construct *clean100*, which consists of the first 100 trains (for each strata, class and problem) of *clean1000*. Since real world data is hardly clean, we also create noisy versions. For each problem, we randomly swap the target class of 5% of our instances, creating the *noisy1000* and *noisy100* datasets.

We end up with 30 simulations for each scenario, noise level, size and method combination. Table 1 reports the AUC-PR mean and standard deviation of each experimental block. When using the *clean* sets, we don't allow any negative examples to be covered by an acceptable clause. When using the *noisy* sets, we allow a negative rule cover of up to 10% of the number of *red* trains.

We compare two methods by using a paired Mann-Whitney test on all their corresponding experiments. Our results show that MF outperforms BASE on all testbeds (p-value = 0.00048). BASE outperforms DPS on size 100 sets (p-value = 0.019), while DPS outperforms BASE on size 1000 (p-value = 0.01). On large noisy sets, DPS outperforms both BASE (p-value = 0.0018) and MF (p-value = 0.0374).



(a) Color *red* trains, specific rule (jagged-roof) in bold



(b) Color *blue* trains, specific-rule (double-hulled) in bold

Fig. 3. A 2-strata 2-class Michalski-train problem

Table 1. AUC-PR mean and standard deviation for each scenario, noise level, size and method combination. Each experimental block is composed of 30 experiments.

Dataset	clean100			clean1000			noisy100			noisy1000		
Method	BASE	MF	DPS	BASE	MF	DPS	BASE	MF	DPS	BASE	MF	DPS
One target rule scenario												
Mean	0.73	**0.83**	0.62	0.87	**0.90**	0.88	0.57	**0.62**	0.54	0.63	0.80	**0.87**
Std dev	0.45	0.34	0.40	0.35	0.24	0.29	0.50	0.47	0.42	0.49	0.36	0.31
Multiple target rules scenario												
Mean	0.61	**0.70**	0.42	0.75	**0.86**	0.77	0.38	**0.52**	0.31	0.52	0.55	**0.65**
Std dev	0.33	0.28	0.29	0.33	0.24	0.30	0.37	0.28	0.32	0.39	0.27	0.29

6.1 Discussion

As one expects, performance improves with larger sets of training examples, and decreases with multiple target rules and noisy sets. The *noisy* runs are harder for three reasons. First is the noise effect *per se*, randomly assigning the wrong target class to 5% of the trains. Second is the 10% minimum positive cover threshold per rule. If a target rule originally narrowly passed this threshold, the addition of noise may decrease its positive coverage below the threshold, and the rule becomes undetectable. Third is the maximum negative cover threshold: in

clean runs, we only consider rules that don't cover any westbound train, which drastically reduces the number of evaluated rules. In *noisy* runs, we allow up to 10% of negative cover. Even if no noise is injected, the exponential expansion of the search space increases the probability that some non-target rule scores better than a target.

It is interesting to note that DPS is the least affected by noise. In each experimental block, DPS suffers the least decrease in mean AUC-PR, none being significant. In the one-target rule and large-set block, adding noise decreases DPS mean by just 1 point, from 0.88 to 0.87 (p-value = 0.94). On the other hand, MF and BASE drop by 10 and 24 percentage points (Table 1). In the four sets of experiments where noise is a variable, DPS drops an average of 8 percentage points, compared to 21.5 for BASE and 20 for MF.

Similarly, DPS improves the most with increasing sample size. In each of the four sets of experiments where size is a variable, DPS displays the highest increase in mean AUC-PR, all of which are significant. In these experiments, DPS increases an average of 32 percentage points, compared to 12 for BASE and 11 for MF (Table 1).

Although no clear pattern emerges from comparing different methods on both one-target and multiple-target scenarios, DPS seems to be slightly more sensitive to the number of target rules. DPS suffers an average decrease of 19 AUC-PR percentage points over the four experimental blocks where target rule scenario is a variable, compared with 13.5 for BASE and 13 for MF (Table 1). Nevertheless, this performance decrease does not alter the method ranking over each experimental block.

In summary, our experiments show that MF is more suitable for either clean data or small datasets. But for large and noisy data, which is what most real world applications are, DPS is more appropriate. In addition, DPS performance increases at a faster rate than MF, and thus may outperform MF for larger clean datasets. DPS, by navigating the differential prediction search space, requires more training examples and generates a set of rules as a consistent theory which explains the data. In contrast, MF and BASE select individual rules that may be suboptimal.

7 Breast Cancer Diagnosis

Our motivating application is to learn older-specific in situ breast cancer differential predictive rules. We apply our three methods to the breast cancer data used in [25]. The data consists of two cohorts: patients younger than 50 years old form the *younger* cohort, while patients aged 65 and above form the *older* cohort. The older cohort has 132 in situ and 401 invasive cases, while the younger one has 110 in situ and 264 invasive.

The data is organized in 20 extensional relations that describe the mammogram, and 35 intensional relations that connect a mammogram with related mammograms, discovered at the same or in prior visits. The background knowledge also maintains information on prior surgeries.

We use the same experimental setting as for the synthetic data, but set $nodes = 200,000$ since the number of predicates is much larger. The BASE method does not return any rule, which highlights the difficulty of this task. Lacking ground truth, we use uplift curves to compare MF and DPS (Fig. 4). DPS consistently outperforms MF, which in turn consistently outperforms a baseline random classifier. DPS has an area under the curve (taken to the baseline) of 16.5, almost double the 9.1 of MF.

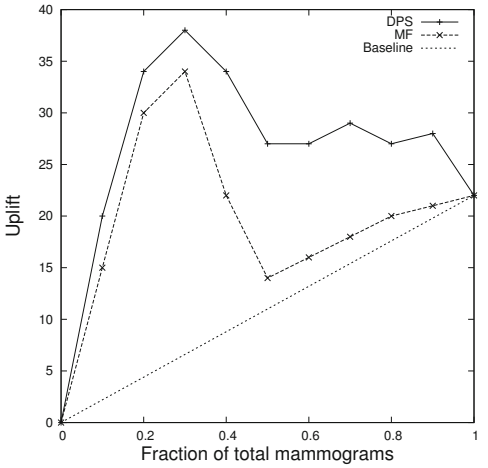


Fig. 4. Uplift curve for breast cancer stage

MF returns 4 differential predictive rules that have a significantly better precision and recall [16] over the older cohort. DPS returns 15. A practicing radiologist, fellowship-trained in breast imaging, examined and assessed all the rules. One MF rule was not found meaningful, while the remaining three are redundant to each other and translate to:

1. Tumor is older-specific in situ if its principal mammographic finding is calcification or single dilated duct, and patient does not have prior surgery.

Single dilated duct is a rare finding and was combined with calcification in our data for convenience. Based on this rule, the more common finding, calcification, is a differential predictor of in situ disease in older patients, which is a novel and interesting result. A possible explanation is that, in asymptomatic women, in situ disease is often associated with screen-detected micro-calcifications; while in symptomatic women, in situ is associated with a palpable mass or pathological nipple discharge [26]. Younger women tend to have more rapidly proliferating cancers that develop into a palpable mass [15], in contrast to more indolent, non-palpable in situ disease manifest as micro-calcification in older patients. This previously unreported finding merits further investigation.

DPS provides a more complete picture of older-specific in situ differential predictors. All 15 returned rules are meaningful and, in addition to extracting the rule described above, four additional themes emerge. DPS is thus able to detect more differentially predictive features than MF, offering a better insight into the medical problem. We select representative clauses from each theme. Tumor is older-specific in situ if:

2. Patient had prior in situ biopsy, and examined-breast had a BI-RADS score of 1 during a previous mammogram, which was not the first visit.
3. Patient had prior in situ biopsy, its examined-breast BI-RADS increased by at least 3 since a previous visit, whereas its other-breast BI-RADS remained constant.
4. Principal mammographic finding is calcification or single dilated duct, examined-breast BI-RADS score increased by at least 3 since a previous visit, and patient had an even earlier screening mammogram.
5. Patient has a breast density of 2, is having a unilateral exam, doesn't have a focal asymmetric density, and principal mammographic finding is calcification or single dilated duct.

Besides calcification, the second DPS rules theme is the presence of a prior in situ biopsy (rules 2, 3). A prior history of biopsy revealing in situ disease is thus a better predictor of in situ recurrence in older women. This observation is partially explained by the longer life span of older women which offers more time for a recurrence to manifest. But this rule may also relate to the indolent nature of in situ breast cancer in older women. In fact, both invasive and in situ tumors in older patients tend to be less aggressive and have lower rates of local recurrence than tumors in younger patients [15]. More specifically, younger women with in situ disease are more likely to progress to an invasive recurrence rather than develop another in situ tumor when they recur [35].

The third theme is the increase in the examined breast BI-RADS score (rules 3, 4). The BI-RADS score is a number that summarizes the examining radiologist's opinion and findings concerning the mammogram [4]. The radiologist assigns a score for each examined breast. An increase in the BI-RADS score over multiple visits reflects increasing suspicion of malignancy. This may be a more pronounced feature in older women because they have more prior mammograms.

The next observation, whereas screening visits predict in situ in older women (rule 4), may also relate to the greater opportunity for screening in older patients. Regular screening mammography is usually recommended for women aged 40 and above. Younger women are more likely to seek care for a palpable lump detection rather than via screening [15]. Thus older women tend to have more screening exams because of regular visits after age 40.

Finally we note a class 2 breast density, out of an increasing density scale of 1 to 4 (rule 5). This is a relatively low breast density, more common in older women, since breast density decreases with age [18]. This rule is of special relevance since it doesn't link to any previous mammogram or history predicate, hence leveling the playing field between younger and older in terms of time. It requires a class

2 breast density and an observed calcification during a unilateral (and hence diagnostic) exam. A lower breast density significantly increases mammogram sensitivity [22], allowing for easier micro-calcification detection.

8 Future Work

This work can be extended in several directions. First, our differential prediction search can be tested and validated using a larger experimental set. We can systematically vary the sample size to establish a performance-size curve, and try different scoring functions. We can also fine grain the construction of the Michalski-trains sets by monitoring the coverage of each target or common rule. Noting that we defined westbound as not-eastbound, it would be interesting to gauge model differences if westbound was defined using a separate set of rules.

Second, this work assumes the presence of a stratified dataset. Given a non-stratified dataset, we may be able to select the best dividing attribute that maximizes differential predictive rules performance. We can repeatedly stratify the data using each of its attributes, and perform differential prediction. We then select the stratification achieving the best results. This approach may be used for differential subgroup discovery.

Third, we only proposed solutions for the 2-strata 2-class differential prediction problem. We plan on extending it to multi-strata problems using f-divergence functions. This being the first attempt at relational differential prediction, we can similarly extend our approach to decision-tree learners.

9 Conclusion

In this work, we extend differential prediction to the multi-relational domain using ILP. We devise and implement three methods to learn 2-strata 2-class differential predictive rules. The first baseline method merges the two strata together while including the stratifying attribute as an additional predicate. The model filtering method generates rules on the target stratum and tests them for differential prediction on the other stratum. The differential prediction search approach alters the ILP search space to use a differential-prediction-sensitive scoring function to assess rules over both strata during rule construction. Our experiments over synthetic data show that the model filtering method is more suitable for either clean or small datasets. For large and noisy data, which is what most real world applications are, the differential prediction search method outperforms both the baseline (p-value = 0.0018) and the model filtering (p-value = 0.0374) approaches. We apply our methods on a breast cancer dataset, and extract novel rules linking calcification to *in situ* disease in older women.

Acknowledgment. This work is supported by US National Institute of Health (NIH) grant R01-CA127379-01. We thank Kendrick Boyd for his help in computing AUC-PR. VSC was funded by the ERDF through the Progr. COMPETE, the Portuguese Gov. through FCT, proj. HORUS ref. PTDC/EIA-EIA/100897/2008,

ADE (PTDC/ EIA-EIA/121686/2010), and the EU Sev. Fram. Progr. FP7/2007-2013 under grant aggrm. 288147.

References

1. Allegra, C.J., Aberle, D.R., Ganschow, P., Hahn, S.M., Lee, C.N., Millon-Underwood, S., Pike, M.C., Reed, S., Saftlas, A.F., Scarvalone, S.A., Schwartz, A.M., Slomski, C., Yothers, G., Zon, R.: National Institutes of Health State-of-the-Science Conference Statement: Diagnosis and Management of Ductal Carcinoma In Situ. J. Natl. Cancer Inst. 102(3), 161–169 (2010)
2. American Cancer Society: Breast Cancer Facts & Figures 2009-2010. American Cancer Society, Atlanta, USA (2009)
3. American Cancer Society: Cancer Facts & Figures 2009. American Cancer Society, Atlanta, USA (2009)
4. American College of Radiology, Reston, VA, USA: Breast Imaging Reporting and Data System (BI-RADS^{TM}), 3rd edn. (1998)
5. American Educational Research Association/American Psychological Association/National Council on Measurement in Education: The Standards for Educational and Psychological Testing (1999)
6. Blockeel, H., De Raedt, L.: Top-down induction of first-order logical decision trees. Artif. Intell. 101(1-2), 285–297 (1998)
7. Chyou, P.H.: Patterns of bias due to differential misclassification by case control status in a case control study. European Journal of Epidemiology 22, 7–17 (2007)
8. Cleary, T.A.: Test bias: Prediction of grades of negro and white students in integrated colleges. Journal of Educational Measurement 5(2), 115–124 (1968)
9. Davis, J., Burnside, E.S., de Castro Dutra, I., Page, D., Ramakrishnan, R., Santos Costa, V., Shavlik, J.: View Learning for Statistical Relational Learning: With an application to mammography. In: Proceedings of the 19th International Joint Conference on Artificial Intelligence, Edinburgh, Scotland, pp. 677–683 (2005)
10. Davis, J., Goadrich, M.: The relationship between precision-recall and ROC curves. In: Proc. of the 23rd International Conference on Machine Learning, Pittsburgh, PA, pp. 233–240 (2006)
11. De Raedt, L.: Logical and Relational Learning. Springer (2008)
12. Demšar, J.: Statistical comparisons of classifiers over multiple data sets. Journal of Machine Learning Research 7, 1–30 (2006)
13. Flegal, K.M., Keyl, P.M., Nieto, F.J.: Differential misclassification arising from nondifferential errors in exposure measurement. American Journal of Epidemiology 134(10), 1233–1244 (1991)
14. Friedman, N., Geiger, D., Goldszmidt, M.: Bayesian network classifiers. Machine Learning 29, 131–163 (1997)
15. Gajdos, C., Tartter, P.I., Bleiweiss, I.J., Bodian, C., Brower, S.T.: Stage 0 to stage III breast cancer in young women. J. Am. Coll. Surg. 190(5), 523–529 (2000)
16. Goutte, C., Gaussier, E.: A probabilistic interpretation of precision, recall and F-score, with implication for evaluation. In: Proc. of the 27th European Conference on IR Research, pp. 345–359. Santiago de Compostela, Spain (2005)
17. Hansotia, B., Rukstales, B.: Incremental value modeling. Journal of Interactive Marketing 16(3), 35–46 (2002)
18. Kelemen, L.E., Pankratz, V.S., Sellers, T.A., Brandt, K.R., Wang, A., Janney, C., Fredericksen, Z.S., Cerhan, J.R., Vachon, C.M.: Age-specific trends in mammographic density. American Journal of Epidemiology 167(9), 1027–1036 (2008)

19. Larson, J., Michalski, R.S.: Inductive inference of VL decision rules. ACM SIGART Bulletin 63, 38–44 (1977)
20. Linn, R.L.: Single-group validity, differential validity, and differential prediction. Journal of Applied Psychology 63, 507–512 (1978)
21. Lo, V.S.: The true lift model - a novel data mining approach to response modeling in database marketing. SIGKDD Explorations 4(2), 78–86 (2002)
22. Mandelson, M.T., Oestreicher, N., Porter, P.L., White, D., Finder, C.A., Taplin, S.H., White, E.: Breast density as a predictor of mammographic detection: comparison of interval- and screen-detected cancers. J. Natl. Cancer Inst. 92(13), 1081–1087 (2000)
23. Muggleton, S.: Inductive Logic Programming. New Generation Computing 8(4), 295–318 (1991)
24. Muggleton, S.: Random train generator (1998), http://www.doc.ic.ac.uk/textasciitildeshm/Software/GenerateTrains/
25. Nassif, H., Page, D., Ayvaci, M., Shavlik, J., Burnside, E.S.: Uncovering age-specific invasive and DCIS breast cancer rules using Inductive Logic Programming. In: 1st ACM International Health Informatics Symposium, Arlington, VA, pp. 76–82 (2010)
26. Patani, N., Cutuli, B., Mokbel, K.: Current management of DCIS: a review. Breast Cancer Res. Treat 111(1), 1–10 (2008)
27. Radcliffe, N.J., Surry, P.D.: Differential response analysis: Modeling true response by isolating the effect of a single action. In: Credit Scoring and Credit Control VI, Edinburgh, Scotland (1999)
28. Radcliffe, N.J., Surry, P.D.: Real-world uplift modelling with significance-based uplift trees. White Paper TR-2011-1, Stochastic Solutions (2011)
29. Rzepakowski, P., Jaroszewicz, S.: Decision trees for uplift modeling. In: 2010 IEEE International Conference on Data Mining, Sydney, Australia, pp. 441–450 (2010)
30. Sackett, P.R., Laczo, R.M., Lippe, Z.P.: Differential prediction and the use of multiple predictors: The omitted variables problem. Journal of Applied Psychology 88(6), 1046–1056 (2003)
31. Sahami, M., Dumais, S., Heckerman, D., Horvitz, E.: A bayesian approach to filtering junk e-mail. In: AAAI Workshop on Learning for Text Categorization, Madison, WI (1998)
32. Santos Costa, V.: The life of a logic programming system. In: de la Banda, M.G., Pontelli, E. (eds.) Proceedings of the 24th International Conference on Logic Programming, Udine, Italy, pp. 1–6 (2008)
33. Society for Industrial and Organizational Psychology: Principles for the Validation and Use of Personnel Selection Procedures, 4th edn (2003)
34. Srinivasan, A.: The Aleph Manual, 4th edn. (2007), http://www.comlab.ox.ac.uk/activities/machinelearning/Aleph/aleph.html
35. Vicini, F.A., Recht, A.: Age at diagnosis and outcome for women with ductal carcinoma-in-situ of the breast: A critical review of the literature. Journal of Clinical Oncology 20(11), 2736–2744 (2002)
36. Young, J.W.: Differential validity, differential prediction, and college admissions testing: A comprehensive review and analysis. Research Report 2001-6, The College Board, New York (2001)
37. Železný, F., Lavrac, N.: Propositionalization-based relational subgroup discovery with rsd. Machine Learning 62(1-2), 33–63 (2006)