DIFFERENTIAL RELATIONAL LEARNING

By

Houssam Nassif

A dissertation submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy

(Computer Sciences)

at the
UNIVERSITY OF WISCONSIN–MADISON
2012

Date of final oral examination: 08/03/2012

The dissertation is approved by the following members of the Final Oral Committee:

David Page, Professor, Computer Sciences

Jude Shavlik, Professor, Computer Sciences

Jerry Zhu, Associate Professor, Computer Sciences

Elizabeth Burnside, Associate Professor, Radiology

Vítor Santos Costa, Assistant Professor, Biomedical Informatics

To the wonderful women in my life,

Carole, Ghia and Yasma Nai.

ACKNOWLEDGMENTS

I would like to thank the many people who have been instrumental in the completion of my thesis.

First are my parents, Georges and Najat, who made my Wisconsin experience possible, from providing me with a cultural environment and fostering the desire of learning in me since I was a child, to their never-ending love and support. My siblings, Wael and Rawane, for all our shared memories and experiences, and the hours spent at the other end of the line especially during tense times. The Mouawad family, Samir, Nawal, Raja and Elie who embraced me as one of their own.

My mentors and advisers, who guided me through my graduate years. Jude Shavlik and Olvi Mangasarian during my first year, and David Page for his valuable advice over the following five years. David provided a flexible research environment encouraging the pursuit of novel ideas, which made this work possible and fostered my growth as a researcher. Elizabeth Burnside, the grant primary investigator, for her constant motivation and patient explanation of everything related to mammography. My prelim and thesis defense committee, David Page, Jude Shavlik, Elizabeth Burnside, Jerry Zhu and Vítor Santos Costa.

Perry Kivolowitz, for going out of his way and granting me a life-saving TA-ship as a new non-guaranteed funding student, as I was fleeing war-ravaged Lebanon during the 2006 Lebanon War, going through four other countries before reaching the US. The always-smiling Angela Thorp, who helped me navigate through the various university requirements.

My research collaborators for the great fun we had working together. Especially: Walid Keirouz my former adviser, Sawsan Khuri and Hassan Al-Ali, we started as an American University of Beirut bioinformatics team, and fanned each to a different institution. Jose

Santos and Bethany Percha for two successful journal papers without ever meeting face-to-face. Vítor Santos Costa and Inês Dutra for invaluable Prolog help and a constant exchange of ideas. Yirong Wu, Ryan Wood, Mehmet Ayvaci and Jagpreet Chhatawal for a productive mammography collaboration.

My office and lab mates, for the hours we spent talking, thinking and joking. Louis Oliphant, Burr Settles, Tim Chang, Jeremy Weiss, Angela Dassow, Trevor Walker, Aubrey Barnard, Yadi Ma, Daniel Wong, Natasha Eilbert, Jie Liu, Kendrick Boyd, Finn Kuusisto, Eric Lantz, Jesse Davis, Bess Berg, Steve Jackson, and many others.

The grant that funded my work, the US National Institute of Health (NIH) grant R01-CA127379-01.

My family away from home, Mike and MaryPat Feifarek, Benny and Jenny Iskandar, and Rob and Holly Kulow. They stood with us during our deepest losses, toughest times, and helped celebrate our best moments. Even though we have no relatives in the US, we never felt alone. They were always there for us. Our stay in Madison and the US would have been drastically different without them.

All my friends who made Madison be the special place it is. Lars Grabow and Arun Rao for introducing me to the city and to their friends since day one. Paul Eppers and Rahul Nabar for providing long-term accommodation and transportation. My neighbors Dave Benzschawel and Chris Bootz, for all the adventures we had in the wild together. Slow Food for a great culinary experience. Hoofers Outing Club, with whom I navigated the river ways of the state. The Wisconsin Speleological Society, especially Dave Wysocki and Dan Pertzborn, with whom I explored upper-Midwest caves. Joe Senulis (N9TWA) who got me into bat monitoring and amateur radio. First Sergeant Lyle Laufenberg and my civil war reenactment unit (4th US Light Artillery, Battery B) for providing a living history window, and teaching mid-19th century military skills.

My closest friends, Matt Feifarek, Edmond Ramly and Torrey Kulow, for being themselves. And for all the time we spent living together, engaging in intellectual discussions,

planning and doing road trips, cooking and dining, watching movies, enjoying our times and being there for each other.

Most importantly is my wonderful Carole, my best encounter ever. Her tender love and intelligent support have no boundaries. I enjoy every minute we spend together, miss her every minute we are apart, and am still discovering more reasons to love her. After a very short journey with our little Ghia, Carole and I are three again. Yasma Nai, for the delight of giving you your night-shift feedings, changing your diapers, and watching you grow. We will so enjoy getting to know each other...

Thank you all for being part of this leg of my journey.

TABLE OF CONTENTS

Page

Appendix

Page

Appendix

Page

LIST OF TABLES

Appendix
Table Page

Appendix
Table

LIST OF FIGURES

DIFFERENTIAL RELATIONAL LEARNING

Houssam Nassif

Under the supervision of Professor David Page

At the University of Wisconsin-Madison

Differential prediction is defined as the case where the best prediction equations and/or the standard errors of estimate are significantly different for different groups of examinees. Maximizing the differential prediction over specific data subsets is an interesting research problem with several real-world applications. This work represents the first attempt to address the multi-relational differential prediction problem. Our approaches are based on Inductive Logic Programming (ILP), which we use to learn differential rules.

We explore several differential methods for learning differential rules in a two-class two-strata system. First we propose the *Model Filtering* (MF) approach, which builds a rule model on the target stratum, and then selects rules that exhibit a differential performance on the other stratum. Second we propose the *Differential Prediction Search* (DPS) method, which alters the search space to consider both strata while scoring rules according to their differential prediction score. Unlike the first two automated methods, the third approach, *Expert Driven* (ED), builds a model on each dataset and lets an expert compare them and infer differential rules.

We compare these methods over a synthetic dataset, and over two important biomedical applications: modeling hexose-protein binding sites, and identifying age-specific breast cancer stage rules. In doing so, we devise the first glucose-binding classifier, empirically validate biochemical hexose-binding knowledge, report the first instance of differential predictive rules discovery, and infer new hexose-binding and breast-cancer dependencies.

Our results show that, for large and noisy data, which is what most real world applications are, DPS is more appropriate. For small and non-noisy data, MF outperforms DPS. We also

augment a Bayes Net with differential rules for risk prediction, and observe a significant performance increase.

Finally, two off-shoots emerged from the main line of work. First, we alter the recall selection of the ILP system ProGolem, establishing that randomized-recall ProGolem should be used by default. Second, we present an information extraction method for free-text mammogram reports, resulting in the first successful mammography information extraction application. We also confirm the application of this method on another dataset and in another language, namely creating the first Portuguese mammography information extraction application.

David Page

Chapter 1

Introduction

Classical classification problems focus on segregating between two or more target classes, by maximizing a given statistic (e.g., accuracy, area under the precision-recall curve). Nevertheless, the predictive power of a classifier can vary across the input space; the classifier may exhibit significant differences in performance over particular instance subgroups. Capturing and modeling this *differential prediction* allows for a deeper understanding of the underlying problem, context-specific decision making, and identification of diverging data subsets.

Building classifiers sensitive to differential prediction is an open research field, and can be seen as a second-order classification problem. Differential prediction often arises as a by-product of standard machine learning problems. A classifier is trained on a dataset, and it may or may not have differential prediction with respect to certain subgroups. An interesting research problem is to construct a classifier that maximizes differential prediction over specific data subsets. This task often arises in the context of analysis of relational databases consisting of multiple tables or relations, known as *multi-relational* data sets. We here present *the first work that explores approaches to address the multi-relational differential prediction problem*. Our approaches are based on Inductive Logic Programming (ILP), and we evaluate them in the context of discovery in two biomedical domains.

1.1 Differential Prediction

A recurrent problem in social sciences is to understand why two or more different populations exhibit differences in a trait. In psychology [29, 72, 137], one may want to assess

the fairness of a test over several different populations. In marketing [56, 75, 103], one may want to compare subjects and controls in order to study the effectiveness of an advertising campaign. Similar tasks arise in other domains and, depending on the domain, the problem is known as *differential prediction* [137], *differential response analysis* [103], or *uplift modeling* [104].

Originally used by psychologists to assess the fairness of cognitive and educational tests, differential prediction is defined as the case where the best prediction equations and/or the standard errors of estimate are significantly different for different groups of examinees [137]. Initially assessed using linear regression, differential prediction arises when a common regression equation results in systematic nonzero errors of prediction for subgroups. This phenomenon is detected by fitting a regression model for each subgroup, and comparing the resulting models [29, 72].

An example is assessing how SAT test scores predict first year cumulative GPA 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.

In contrast to most studies of *differential prediction* in psychology, marketing's *uplift modeling* assumes an active agent. It directly models the incremental impact of a treatment, such as a direct marketing action, on the behavior of a set of individuals. The SAT score doesn't actively change GPA, whereas a marketing action does actively change behavior. In both cases, the population is stratified into predefined sub-populations (henceforth called strata), and we aim at detecting and modeling the class differential prediction over the stratified data. We thus argue that the concepts and techniques originally developed for uplift marketing can, and should, apply to the task of differential prediction (and vice versa).

Starting from a one-variable simple regression, differential prediction has been studied extensively in the context of multi-attribute data [104, 112]. One approach is to generate different classifiers for each given subgroup, and to look for the main differences between the

classifiers, as typically done in psychology. Further progress requires building models driven by differential evaluation functions [111].

1.2 Thesis Statement

My thesis is that *ILP-based differential relational classifiers can effectively propose rules that apply to a given multi-relational data subset, maximize performance differences over a stratified dataset, and offer significant insight into the underlying domain.* My work is motivated by two biomedical applications: modeling hexose-protein binding sites, and identifying age-specific breast cancer stage rules.

Even though our work obeys the main postulates followed by prior work in uplift modeling [111], we observe that, to the best of our knowledge, this is 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. Our work focuses on understanding factors that describe differential performance.

In this work, we explore several differential methods for learning differential rules in a two-class two-strata system. A very basic method is the *Expert Driven* (ED) approach, which builds a model on each dataset, and lets an expert compare the two and infer differential rules. A fully automated method is the *Model Filtering* (MF) approach, which builds a rule model on the target stratum, and then selects rules that exhibit a differential performance on the other stratum. The third approach is the *Differential Prediction Search* (DPS) method, which alters the search space to consider both strata while scoring rules according to their differential prediction score.

1.3 ILP for Differential Prediction

ILP is a machine learning approach that learns a hypothesis, composed of a set of rules in first-order logic, that explains a given dataset [79]. In standard classification, ILP has three major advantages over other machine learning and data mining techniques. First, it allows an easy interaction between humans and computers by using background knowledge

to construct hypotheses and guide the search. Second, it returns results in an easy-to-understand if-then format. Finally, it can operate on data in a relational database, because such databases are a theoretical subset of first-order logic.

In the context of differential prediction, ILP — as a rule-learning technique — has a fourth major advantage. 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 particular rules. 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. These rules are subgroup-specific due to their differential predictive ability. We are not aware of any prior use of rule-learners to identify differential predictive rules.

One aim of this work is to *formally define the differential predictive rule identification paradigm*. Another is to *implement it within the ILP framework*. A third is to *apply it to important biomedical domains*.

1.4 Document Overview

The rest of this document is organized as follows. Chapter 2 reviews prior differential prediction work, and formally define the task of learning differential predictive rules. Chapter 3 covers the necessary background, it overviews ILP systems, the datasets we use, and our comparison methodology. Our work is driven by two main applications, identifying age-specific breast cancer stage rules, and modeling hexose-protein binding sites. Chapters 4, 5 and 6 present three different differential predictive rule learning techniques. Chapter 7 is a hexose application off-shoot, where we alter the recall selection method of the ProGolem ILP system. Chapter 8 explains a necessary information extraction preprocessing step for mammography free-text records. Chapter 9 concludes with a summary and future work suggestions.

Chapter 2

Differential Prediction

The problem of differential prediction, where one wants to capture and model differences between two or more subgroups, arises independently in a variety of fields. In this chapter *we review prior work on differential prediction* in greater detail. We close this chapter with *a novel formulation of differential predictive concepts.*

2.1 Regression Usage

Differential prediction was first used in Psychology to assess the fairness of cognitive and educational tests. It is defined as the case where the best prediction equations and/or the standard errors of estimate are significantly different for different groups of examinees [137]. 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 [29, 72]. 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 [6, 119]. 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.

Coming back to the earlier SAT example in Section 1.1, we fit a regression model for each gender group (Figure 2.1). If the slopes or intercepts are significantly different between both models, then the SAT test exhibits differential prediction with respect to gender.

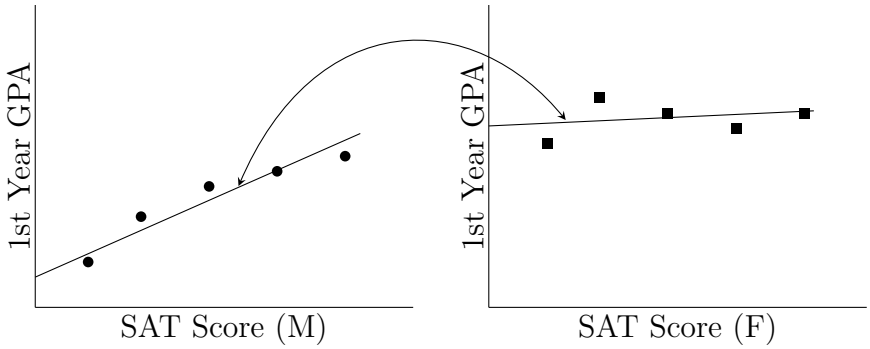


Figure 2.1: Using regression to detect differential prediction. Fit a regression model for each group and compare both models. SAT exhibits differential prediction across gender if the models are significantly different.

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 [27, 43]. This concept is the basis of the related machine learning concept of "differential misclassification cost", incorporating different misclassification costs into a cost sensitive classifier [113]. During the training phase, such a classifier would assign different misclassification costs for various subgroups (usually for each class), and would predict the class with minimum expected misclassification cost.

Examining each predictor separately in a regression analysis may result in a misspecified model. The regression coefficient can be biased if we omit a variable that is related to the target and correlated with a measured predictor variable [73]. This problem is known as the omitted variable problem. It can be leveraged by broadening the selection system to include other relevant predictors in the regression [112].

2.2 Classifier Usage

The classification literature, especially in the medical domain, has extended the differential prediction concept to differences in predicted performance when an instance is classified into one condition rather than into another [119]. Hence differential prediction is detected by comparing the performance of different classifiers on the same subgroup (e.g. [39]), or the same classifier on different subgroups (e.g. [101, 131]).

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 [103], Lo's true lift model [75], and Hansotia and Rukstales' incremental value modeling [56]. 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 [111] 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, the splitting criterion is zero if treatment and control are independent. Third, if the control group is empty, the criterion reduces to a classical splitting criterion. 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 [104] criticize the third postulate and the fact that the measures are independent of population size, a parameter that they consider crucial in practical applications.

2.3 Rules for Differential Prediction

Although, for the best of our knowledge, this work is the first to address differential rule learning, this section reviews other usages of rules for differential prediction.

2.3.1 Indexes of Development

The use of rules to achieve a differential classification is a technique utilized in developmental psychology as a developmental metric to systematically classify linguistic performances into a hierarchical taxonomy of cognitive-structure types [122]. Researchers, through observation and collective informal judgments, identify specific skills that reflect a particular developmental stage [30, 42]. Thus, by inductive and abductive reasoning, researchers manually construct rules — called indexes of development — that classify performances into cognitive types.

Notice that the concept of rule generation and prediction in developmental psychology is different than in machine learning. Rules and indexes are manually created by a panel of experts following observation studies. Dealing with behavioral data, rules are validated according to psychometric validity and reliability parameters [6]; and not according to accuracy or precision. This is often the case in social sciences, where ground truth is typically unknown, and the rule coverage is mainly determined by an expert. The way the resulting rules are viewed as metrics organized in an index of development is closer to a multi-class prediction task, than it is to identifying differential predictive rules.

2.3.2 Relational Subgroup Discovery

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" [138]. 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.

2.3.3 Instance Relabeling

The only other effort we are aware of to identify rules that achieve a differential prediction across a stratified dataset recently came from our research lab. Working on uncovering adverse drug effects, the aim is to find rules covering patient subgroups that have a differential prediction before and after drug administration [96].

They start with an after-drug administration subset with positive P_1 and negative N_1 instances, and a before-drug administration subset with positive P_2 and negative N_2 instances. Using the coverage evaluation function (the number of positives covered by the rule, minus the number of negatives covered), a rule that has a good performance on the target set (after drug administration) and a bad performance on the other set (before drug administration) will result in a high $(cover(P_1) - cover(N_1))$ score and a low $(cover(P_2) - cover(N_2))$ score.

Their methodology consists of redefining the positive set as $(P_1 + N_2)$, and the negative set as $(P_2 + N_1)$, as shown in Figure 2.2. By using the coverage evaluation function, which maximizes the $Pos - Neg$ cover score for a rule, they aim at maximizing:

$$\begin{aligned} Score &= cover(Pos) - cover(Neg) \\ &= cover(P_1 + N_2) - cover(P_2 + N_1) \\ &= (cover(P_1) + cover(N_2)) - (cover(P_2) + cover(N_1)) \\ &= (cover(P_1) - cover(N_1)) - (cover(P_2) - cover(N_2)). \end{aligned} \tag{2.1}$$

Now the differential assumption of a high $(cover(P_1) - cover(N_1))$ score means a high $cover(P_1)$ and a low $cover(N_1)$. The low $(cover(P_2) - cover(N_2))$ score assumption is ambiguous, since it can be fulfilled by a high $cover(P_2)$ and a high $cover(N_2)$, a low $cover(P_2)$ and a low $cover(N_2)$, or a low $cover(P_2)$ and a high $cover(N_2)$.

The Instance Relabeling score does not guarantee returning a rule with high $cover(P_1)$, since a non-differential rule with a high $cover(N_2)$ and low covers for P_1, P_2 and N_1 will

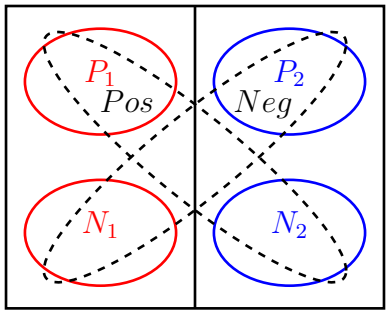


Figure 2.2: Instance relabeling for differential prediction

result in a high score. This scoring function assumes that:

$$\begin{aligned} &\operatorname{argmax} cover(P_1 + N_2) \quad \text{and} \quad \operatorname{argmin} cover(P_2 + N_1) \approx \\ &\operatorname{argmax}(cover(P_1) - cover(N_1)) - (cover(P_2) - cover(N_2)). \end{aligned} \tag{2.2}$$

This is not necessarily the case. The instance-relabeling method may produce suboptimal or even non-differential prediction rules.

It is important to note that this instance-relabeling method is specific to the coverage rule-scoring evaluation function. The instances rearrangement performed in Equation 2.1 doesn't match the differential prediction score of other clause-utility functions. Applying this method to another scoring function necessitates a different instance relabeling schema, or may not be feasible at all.

2.4 Differential Predictive Rule Definition

Given data that can be partitioned into a set of strata, we define a differential predictive concept (expressed as a rule in ILP) 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 2.4.1 (**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 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 examples of D_i with class label l, such that:

$$(\forall (i,j) \in [1,n],\ i \neq j)\ D_i \subset D,\ D_j \subset D,\ D_i \cap D_j = \emptyset, \tag{2.3}$$

$$\forall (i,l)\ D_i^l \neq \emptyset. \tag{2.4}$$

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:

$$\mathcal{D} = \{D_i \mid 1 \leq i \leq k\}. \tag{2.5}$$

As an example, and in the case of our breast cancer application, the concept tc is the cancer stage, with labels l being *invasive* and *insitu*. The set X of instances is the set of mammogram records, and D is the cancer records set. We stratify D according to age, with $k = 3$, resulting in a younger, middle and older strata. In our experiments, we only retain the younger and older subsets, forming a 2-stratified dataset $\mathcal{D}$.

Note that the definition itself can also be relaxed to allow for overlapping groups (not completely stratified). For example, if you stratify by race, some people might belong to multiple races, or if you stratify by the geographical region where people were raised, some might have lived in multiple regions. In this case, we do not require $D_i \cap D_j = \emptyset$.

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

Definition 2.4.2 (**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:

$$\forall i \neq j,\ S(c_j|D_j) \gg S(c_j|D_i). \tag{2.6}$$

The score difference ($\gg$) can be evaluated using statistical significance tests or by comparing against a threshold. In practice, we search a large space of possible concepts/rules for differential predictive ones. In real-world applications, we also want the rules to achieve a minimum level of performance. In this work we focus on 2-strata 2-class differential problems.

Chapter 3

Background

In this chapter, we cover the necessary background before presenting our ILP-based differential prediction methods. This includes an overview of ILP, the datasets we use, and our comparison methodology.

3.1 ILP

Inductive Logic Programming is a class of classifiers that learns rules in first-order logic. There exist multiple ILP systems. In this work, we use two different ILP algorithms, top-down Aleph [121] and bottom-up ProGolem [82].

3.1.1 Terminology

The first-order logic alphabet is composed of *predicate* symbols (e.g. *round*), *function* symbols (e.g. *color*), *constants* (e.g. *Blue*) and *variables* (e.g. x) [79]. Predicates are features that take on true or false values, whereas functions are features that may take constants as their values. Constants are capitalized, while variables are in lowercase.

A *term* is any constant, variable, or function applied to a term (e.g. *Blue*, x, *color(x)*). An *atomic formula* is a predicate symbol together with its arguments, each argument being a term. A *ground atom* (or *fact*) is an atomic formula with no variables (e.g. *sibling(A,B)*). Dataset features are ground atoms, and they constitute the *background knowledge*. A *literal* is an atomic formula or its negation (e.g. *round(x)*, $\neg$ *sibling(x,A)*).

A *clause* is a disjunction of literals whose variables are assumed to be universally quantified [110]. A *Horn clause* is a clause with at most one positive literal, the remaining literals being negated. A *definite clause* is a Horn clause with exactly one positive literal. A definite clause is equivalent to an implication, since

$$H \vee \neg L_1 \vee \neg L_2 \vee \cdots \vee L_n \tag{3.1}$$

can be rewritten as the rule

$$H \leftarrow (L_1 \wedge L_2 \wedge \cdots \wedge L_n), \tag{3.2}$$

$$\text{if } (L_1 \wedge L_2 \wedge \cdots \wedge L_n), \text{ then } H. \tag{3.3}$$

The literal H is the *head* of the clause, while $(L_1 \wedge L_2 \wedge \cdots \wedge L_n)$ constitute the clause *body*.

3.1.2 Bottom Clause

Given a dataset composed of positive and negative instances, an ILP classifier attempts to learn a set of rules (definite clauses) that will correctly discriminate between the two sets. These rules would cover most or all of the positive instances, and little or none of the negative instances.

In the Figure 3.1 example, positive instance A has atomic formulas $red(A)$, $big(A)$, $round(A)$, and $sibling(A, B)$. A bottom clause is the most specific clause that entails the example selected [110]. Since instance A is linked to instance B, the bottom clause of instance A would also include the atomic formulas of instance B.

$$Bottom\ clause(A) : red(A),\ big(A),\ round(A),\ sibling(A, B),\ red(B),\ big(B),\ round(B). \tag{3.4}$$

More formally, given a positive example $pos(x_i)$, let $\perp_i$ be the bottom clause for example i. $\perp_i$ is the most specific hypothesis that, together with the background knowledge B, entails x_i: $(B \wedge \perp_i \wedge x_i) \vdash pos(x_i)$.

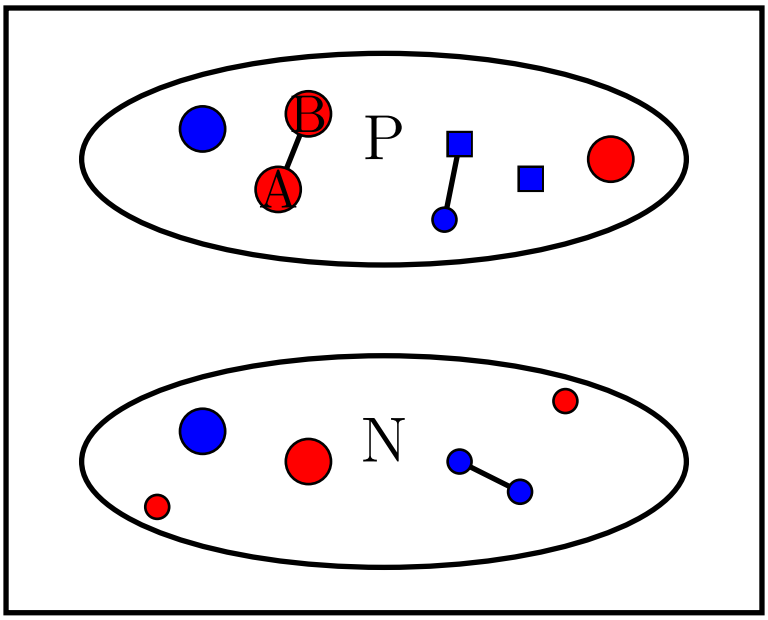


Figure 3.1: ILP example

3.1.3 Aleph

Aleph [121] is an ILP system that implements the Progol algorithm [83]. Progol's main advantage is the use of a bottom clause to guide the search. It randomly selects a positive example $pos(x_i)$ and builds its bottom clause $\perp_i$ during the "saturation" step. The use of a bottom clause ensures that, by construction, all clauses in a refinement graph search are guaranteed to cover at least the example associated with the bottom clause.

Aleph then performs a general-to-specific top-down hypothesis space search, bounded by the most general possible hypothesis and by the bottom clause. To do so, Aleph guides the search using the bottom clause. Starting with the most general hypothesis $pos(\mathbf{X})$, Aleph refines the clause by repeatedly adding literals from the bottom-clause. The new rule will be more specific, covering only a subset of the examples previously covered. This process is the "reduction" step. Algorithm 3.1 highlights the major steps of Aleph.

Pertaining to the Figure 3.1 example, let us suppose instance A is first selected. Aleph would construct $Bottom\ clause(A)$ (Equation 3.4) during the saturation step. It starts its

Algorithm 3.1 Aleph

Require: Examples E, mode declarations M, background knowledge B, Scoring function S

$Learned\_rules \leftarrow \{\}$

$Pos \leftarrow$ all positive examples in E

while Pos **do**

Select example $e \in Pos$

Construct bottom clause $\perp_e$ from e, M and B ▷ Saturation step

$Candidate\_literals \leftarrow Literals(\perp_e)$

$New\_rule \leftarrow pos(\mathbf{X})$ ▷ Most general rule

repeat ▷ Top-down reduction step

$Best\_literal \leftarrow \underset{L \in Candidate\_literals}{\text{argmax}}$ S(New_rule with precondition L)

add $Best\_literal$ to preconditions of New_rule

until No more $S(New\_rule)$ score improvement

$Learned\_rules \leftarrow Learned\_rules + New\_rule$

$Pos \leftarrow Pos -$ {members of Pos covered by New_rule}

end while

return $Learned\_rules$

top-down search by considering any instance to be positive. Aleph then refines this most general hypothesis during the reduction step by adding predicates from *Bottom clause*(A).

3.1.4 ProGolem

ProGolem [82], on the other hand, is a newly developed algorithm combining approaches from Progol [83] and Golem [81]. Like Progol and Aleph, it uses a bottom clause to guide the search. But unlike them, it performs a variant of Golem-like bottom-up search, based on Asymmetric Relative Minimal Generalization (ARMG). It uses ARMG to navigate a specific-to-general bottom-up subsumption order relative to the bottom clause.

Like Aleph, ProGolem randomly selects a positive example $pos(x_i)$ and constructs its bottom clause $\perp_i$ during the saturation step. ProGolem starts the search from this most-specific clause, considering as positives only the examples covered by the bottom clause $\perp_i$. During the reduction step, it successively drops a minimal set of atoms from the body to allow coverage of one additional example. This is done by constructing the ARMG clause of the current clause and the additional example. By dropping this set of literals the clause becomes more general, and will cover a superset of the examples previously covered. Algorithm 3.2 highlights the major steps of ProGolem, it mirrors the Aleph algorithm to highlight search strategy differences.

Algorithm 3.2 ProGolem

Require: Examples E, mode declarations M, background knowledge B, Scoring function S

$Learned\_rules \leftarrow \{\}$
$Pos \leftarrow$ all positive examples in E
while Pos **do**
  Select example $e \in Pos$
  Construct bottom clause $\perp_e$ from e, M and B ▷ Saturation step
  $New\_rule \leftarrow \perp_e$ ▷ Most specific rule
  repeat ▷ Bottom-up reduction step
    Select a different example $e' \in Pos$
    $Blocking\_literals \leftarrow ARMG(New\_rule, e')$
    remove $Blocking\_literals$ from preconditions of New_rule
  until No more $S(New\_rule)$ score improvement
  $Learned\_rules \leftarrow Learned\_rules + New\_rule$
  $Pos \leftarrow Pos - \{$members of Pos covered by $New\_rule\}$
end while
return $Learned\_rules$

Going back to the Figure 3.1 example, let us suppose instance A is first selected. ProGolem too would construct $Bottom\ clause(A)$ (Equation 3.4) during the saturation step.

It starts its bottom-up search by considering as positives only the instances covered by *Bottom clause*(A). ProGolem then refines this most specific hypothesis during the reduction step by dropping selected predicates.

3.1.5 Theory Rules

Both ProGolem and Aleph stop hypothesis refinement when the hypothesis score stops improving. A rule scores well if it covers many positive and few negative examples. If the rule passes a certain performance threshold, it is added to the *theory*, and all the positive examples it covers are removed. The cycle of saturation and reduction continues on the remaining examples. When all positive examples are covered or no new rules can be found, the ILP system outputs its theory, the set of the best rules found so far. Then, in the testing stage, a new instance is classified as positive if it is covered by any of the theory rules, otherwise it is labeled as negative.

In the Figure 3.1 example, a possible theory would be composed of the following rules:

- $P(X)$ if $square(X)$
- $P(X)$ if $red(X) \wedge big(X)$
- $P(X)$ if $sibling(X, Y) \wedge square(Y)$

This theory will result in one false positive (the *red*, *big*, and *round* negative instance) and one false negative (the *blue*, *big*, and *round* positive instance).

Aleph adopts a local theory construction method, incrementally adding a new rule to its theory after each reduction cycle. This method depends on the ordering of the positive examples, and it is possible that the best rules are not generated. This situation may occur if these better rules would be generated by examples that were removed by previous suboptimal rules. By contrast, ProGolem implements a global theory construction approach, which ensures that the theory is only constructed after all rules have been generated. ProGolem repeatedly adds to the theory the rule that best improves the global theory score.

3.2 Mammography Dataset

Our main application is to uncover age-specific breast cancer stage differential prediction rules. We here present our mammography dataset and our preprocessing work to augment it with additional features.

3.2.1 Original Dataset

Our database consists of 146,198 consecutive mammograms recorded at the University of California San Francisco (UCSF) Medical Center between January 6, 1997 and June 29, 2007. In addition to the mammography table, our relational database includes another table consisting of 4,081 biopsies performed between January 7, 1997 and November 18, 2007. Biopsy results are either invasive, in situ or benign. Attempting to discriminate invasive versus in situ cancers based on mammography findings, we identify cancerous biopsies and match each of them with its corresponding diagnostic mammography exam. We end up with 1063 invasive and 412 in situ cancerous diagnostic mammography exam cases.

We separate our data into three cohorts based on age (Table 3.1). We designate patients aged 65 and older as an "older" cohort, patients between 50 and 64 years as a "middle" cohort, and patients less than 50 years old as a "younger" cohort. While we did not stratify by menopausal status; we do know that mean age at menopause among US women is reported to range between 49.1 and 50.5 for different birth cohorts [93]. Therefore, a large proportion of women less than 50 years old (our "younger" cohort) would be premenopausal and a large proportion of women aged 65 and older (our "older" cohort) would be postmenopausal.

Age of subjects ranged from 27 to 97, with mean 59.7 and standard deviation 13.4. The overall proportion of in situ was 27.9% with a slightly higher incidence in our younger and middle groups versus our older group, which is consistent with the literature [40].

3.2.2 Structured and Extracted Features

The mammography reports use a structured format that records patient characteristics and examination findings (Table 3.2). Additional relevant details describing the findings were

Table 3.1: Age-based cohorts

Cohort	Invasive	In Situ	Biopsies	Patients	In Situ (%)
Younger	264	110	374	353	29.4%
Middle	398	170	568	538	29.9%
Older	401	132	533	493	24.8%
Total	1063	412	1475	1384	27.9%

dictated by the interpreting radiologist in free text. Mammography features and findings are based on the American College of Radiology's Breast Imaging Reporting and Data System (BI-RADS) [5]. To extract the BI-RADS descriptors from the dictated text, we developed a Natural Language Processing technique and applied it on the UCSF dataset (Table 3.2). We present our medical text information extraction method in Chapter 8.

Table 3.2: List of structured and extracted features

Structured	Extracted using NLP [91]
Family breast cancer history	Mass margin
Personal breast cancer history	Mass shape
Prior surgery	Calcification distribution
Palpable lump	Calcification morphology
Screening v/s diagnostic	Architectural distortion
Indication for exam	Associated findings
Breast Density	Mammary lymph node
BI-RADS code left	Asymmetric breast tissue
BI-RADS code right	Focal asymmetric density
BI-RADS code combined	Tubular density
Principal finding	Mass size

3.2.3 Extensional Predicates

One of ILP's advantages is that it can naturally operate on a relational database, because such databases are a theoretical subset of first-order logic. To take full advantage of ILP's relational abilities, we extend the basic background knowledge of Table 3.2 by introducing predicates that link related records and features together.

The mammography table schema (Table 3.2) specifies a "left-breast" and a "right-breast" BI-RADS code. A BI-RADS code is a number that summarizes the radiologist's opinion and findings concerning the mammogram [5]. The BI-RADS codes are ranked as $1 < 2 < 3 < 0 < 4 < 5$, in increasing order of malignancy probability. Since we know which breast was biopsied for our target cancerous patients, we convert the left and right BI-RADS features to "this-side" and "other-side" BI-RADS codes. Similarly we change any "left" or "right" value into "this-side" or "other-side".

For example, suppose one of our records is identified by `UniqueID = 21` and has the following features: `FamilyHistory = None`, `BiradsCodeLeft = 4`, `PalpableLump = Right`. By consulting the biopsy table, we find that the left breast was biopsied. We thus convert `BiradsCodeLeft` to `ThisSideBirads = 4`, and the value `Right` to `OtherSide`. The resulting ILP predicates would be: `FamilyHistory(21, None)`, `ThisSideBirads(21, 4)`, `PalpableLump(21, OtherSide)`.

We then extend this basic background knowledge by linking each patient's cancerous diagnostic mammography record to the patient's other previous screening or diagnostic mammograms (Table 3.3). This link allows ILP to access and learn from the patient's previous mammography history. In addition, we add predicates that monitor mass size change and BI-RADS code change when compared to older mammography studies, as well as predicates detecting the occurrence and location of prior biopsies (Table 3.3). This allows ILP to compare mass sizes to given size intervals, and different BI-RADS codes to each other.

Table 3.3: List of ILP extended predicates

first diagnostic mammogram (id)
old study (id, old id)
old biopsy (id, old id, result)
old biopsy same location (id, old id, result)
mass size decrease (id, old id)
mass size increase (id, old id)
this side BI-RADS old study (id, old id, old BI-RADS)
other side BI-RADS old study (id, old id, old BI-RADS)
combined BI-RADS old study (id, old id, old BI-RADS)
this side BI-RADS decrease (id, old id)
other side BI-RADS decrease (id, old id)
combined BI-RADS decrease (id, old id)
this side BI-RADS increase by at least X (id, old id)
other side BI-RADS increase by at least X (id, old id)
combined BI-RADS increase by at least X (id, old id)

3.3 Synthetic Michalski-Trains Dataset

In addition 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 [70] 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:

$$east(T) \text{ :- } infront(T, C1, C2), short(C1), long(C2). \tag{3.5}$$

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

$$east(T) \text{ :- } has\_car(T, C), jagged(C). \tag{3.6}$$

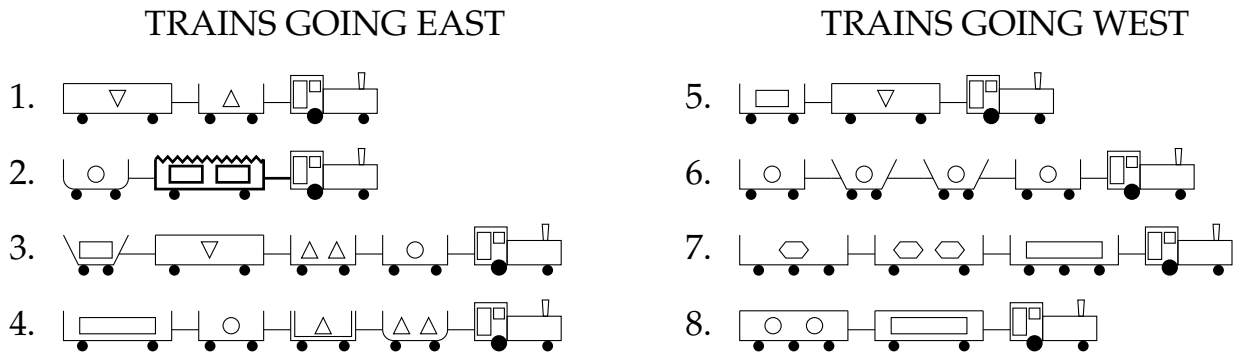
Stratum *blue* specific eastbound rule:

$$east(T) \text{ :- } has\_car(T, C), double(C). \tag{3.7}$$

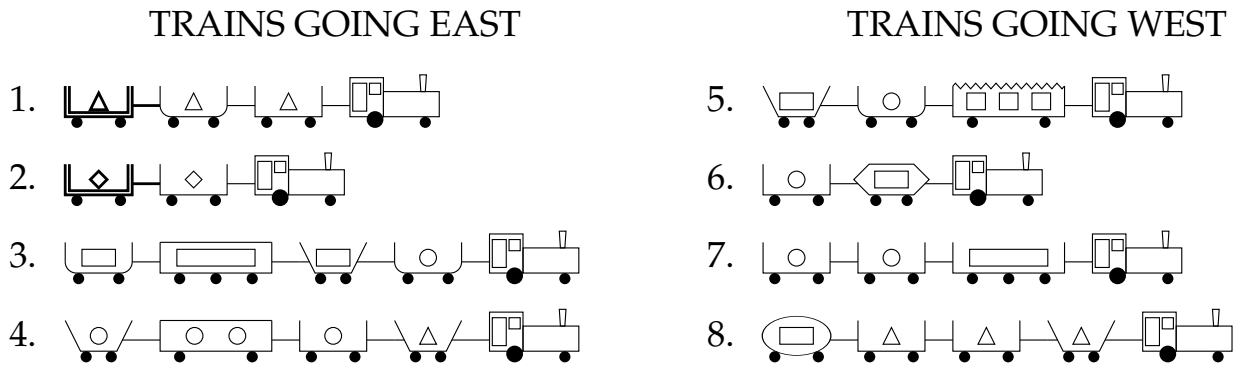
Figure 3.2(a) shows *red* trains, where eastbound trains 1, 3 and 4 have a short carriage in front of a long one (common rule), while train 2 has a jagged roof carriage (*red* specific rule). Figure 3.2(b) shows *blue* trains, where eastbound trains 3 and 4 follow the common rule, while trains 1 and 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 were *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 [80] 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*.



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



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

Figure 3.2: A 2-strata 2-class Michalski-train problem

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. 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 generate 30 simulations for each scenario, noise level and size combination.

3.4 Hexose Dataset

We also consider a secondary application, inferring differences between specific glucose and general hexose binding. We collect a hexose dataset and extract multiple chemical and spatial features from the binding site.

3.4.1 Dataset Collection

Due to the crucial importance of 3-D structure for protein binding, our model should be based on 3-D spatio-chemical data. The Protein Data Bank (PDB) [10] is the largest repository of experimentally determined and hypothetical three-dimensional structures of biological macromolecules. We mine it for proteins crystallized with the most common hexoses: galactose, glucose and mannose [44]. We ignore theoretical structures and files older than PDB format 2.1. We eliminate redundant structures using PISCES [134] with a 30% overall sequence identity cut-off. We use Swiss-PDBViewer [54] to detect and discard sites that are glycosylated or within close proximity to other ligands. We check the literature to ensure that no hexose-binding site also binds non-hexoses. The final outcome is a non-redundant positive data set of 80 protein-hexose binding sites (see Appendix Table A.1). Since, the number of binding-sites crystallized with glucose is small (35), we use the same glucose data as part of our hexose dataset (relaxed version of Definition 2.4.1).

We also extract an equal number of negative examples. The negative set is composed of non-hexose binding sites and of non-binding surface grooves. We choose 22 binding-sites that bind hexose-like ligands: hexose or fructose derivatives, 6-carbon molecules, and molecules similar in shape to hexoses (see Appendix Table A.2). We also select 27 other-ligand binding sites, ligands who are bigger or smaller than hexoses (Table A.2). Finally, we specify 31 non-binding sites: protein surface grooves that look like binding-sites but are not known to bind any ligand (see Appendix Table A.3).

We use 10-fold cross-validation to validate our approach. We divide the data set in 10 stratified folds, thus preserving the proportions of the original set labels and sub-groups.

3.4.2 Binding Site Representation

We view the binding site as a sphere centered at the ligand, as portrayed in Figure 3.3. We subdivide the sphere into concentric layers [9]. We compute the center of the hexose-binding site as the centroid of the coordinates of the hexose pyranose ring's six atoms. For negative sites, we use the ligand's central point when a ligand is present, and the center of the cavity when a ligand is missing. The farthest pyranose-ring atom from the ring's centroid is located 2.9 Å away. Considering atomic interactions to be significant within a 7 Å range [12], we fix the binding site sphere radius to 10 Å. Given the molecule and the binding site centroid, we extract all protein atoms within the sphere, as well as water molecules and ions present in the binding groove. We discard hydrogen atoms since most PDB entries lack them.

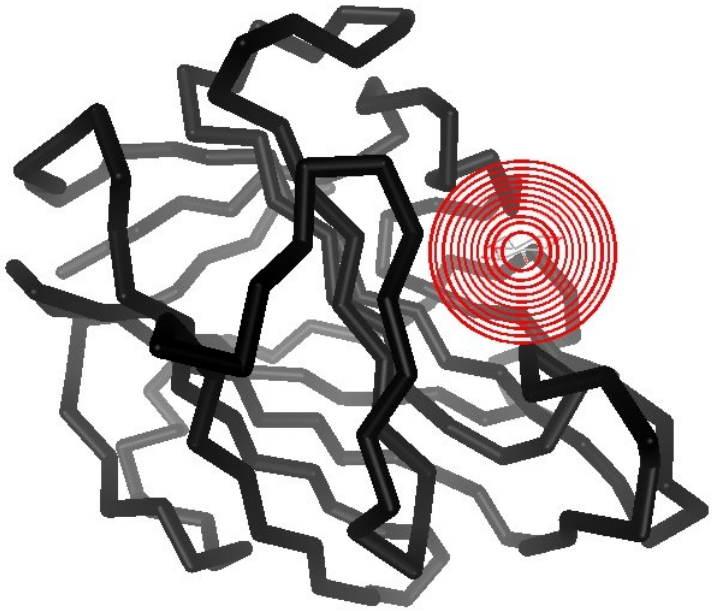


Figure 3.3: Glucose bound to a hydrolase, PDB entry 1I8A. The concentric-layers binding site center is the centroid of the glucose pyranose ring.

For every extracted atom we record its PDB-coordinates, its charge, hydrogen bonding, and hydrophobicity properties, the residue group it belongs to, and its atomic element and name. Every PDB file has orthogonal coordinates and all atom positions are recorded accordingly. The partial charge measure per atom is positive, neutral, or negative; atoms can

form hydrogen bonds or not; hydrophobicity atomic measures are considered as hydrophobic, hydroneutral, or hydrophilic. Amino acids are generally categorized into subgroups, based on the structural and chemical properties of their side chains [11, 125]. Finally, every PDB-atom has an atomic element and a specific name. Tables 3.4 and 3.5 detail the amino-acid grouping and the atomic feature values used.

Table 3.4: Residue subgrouping

Category	Residues
Aromatic	Phe, Tyr, Trp, His
Aliphatic	Ala, Val, Leu, Ile, Met
Neutral	Gln, Asn, Ser, Thr, Pro, Gly, Cys
Acidic-carboxylate	Glu, Asp
Basic	Lys, Arg

3.5 Comparing Differential Prediction Results

When using synthetic data to uncover differential predictive rules, we know the ground truth. We thus 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 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 [34]. 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 [35]. We set the confidence level to 95%.

Table 3.5: Chemical atomic features. Charge is either positive, neutral or negative. Atoms are either capable of forming hydrogen bonds, or are not. Hydrophobicity levels are +1 (hydrophobic), 0 (hydroneutral) and -1 (hydrophilic).

Atom Type	Functional Group	Location	Residue	PDB Atom Symbol	Chrg	Hydrophob	H Bond
Oxygen	Amide peptide linkage	Backbone	All	O	0	-1	H Bond
Oxygen	Carboxyl – C terminus	Backbone	All	OXT	-ve	-1	H Bond
Oxygen	Carboxyl	Side Chain	GLU	OE1	-ve	-1	H Bond
Oxygen	Carboxyl	Side Chain	GLU	OE2	-ve	-1	H Bond
Oxygen	Carboxyl	Side Chain	ASP	OD1	-ve	-1	H Bond
Oxygen	Carboxyl	Side Chain	ASP	OD2	-ve	-1	H Bond
Oxygen	Amide	Side Chain	GLN	OE1	0	-1	H Bond
Oxygen	Amide	Side Chain	ASN	OD1	0	-1	H Bond
Oxygen	Hydroxyl	Side Chain	SER	OG	0	-1	H Bond
Oxygen	Hydroxyl	Side Chain	THR	OG1	0	-1	H Bond
Oxygen	Hydroxyl - Phenolic	Side Chain	TYR	OH	0	-1	H Bond
Nitrogen	Amide peptide linkage	Backbone	All except PRO	N	0	-1	H Bond
Nitrogen	Amide peptide linkage	Backbone	PRO	N	0	-1	--
Nitrogen	Amide	Side Chain	GLN	NE2	0	-1	H Bond
Nitrogen	Amide	Side Chain	ASN	ND2	0	-1	H Bond
Nitrogen	Amine	Side Chain	LYS	NZ	+ve	-1	H Bond
Nitrogen	Guanidino	Side Chain	ARG	NE	+ve	-1	--
Nitrogen	Guanidino	Side Chain	ARG	NH1	+ve	-1	H Bond
Nitrogen	Guanidino	Side Chain	ARG	NH2	+ve	-1	H Bond
Nitrogen	Imidazole	Side Chain	HIS	ND1	0	-1	--
Nitrogen	Imidazole	Side Chain	HIS	NE2	0	-1	H Bond
Nitrogen	Indole	Side Chain	TRP	NE1	0	0	--
Carbon	Amide peptide linkage	Backbone	All	C	0	0	--
Carbon	C-alpha	Backbone	All	CA	0	0	--
Carbon	Aliphatic – neutral	Side Chain	Set A (See below)	CB, CG, CD, CE	0	0	--
Carbon	Aliphatic – hydrophobic	Side Chain	LEU, VAL, ILE, MET	CB, CG, CD, CE	0	1	--
Carbon	Aliphatic – Branch	Side Chain	LEU, VAL, ILE	CG1, CG2, CD1, CD2, CD1	0	1	--
Carbon	Phenyl - aromatic	Side Chain	PHE, TYR	CG,CD1, CD2, CE1, CE2, CZ	0	1	--
Carbon	Imidazole	Side Chain	HIS	CG, CD2, CE1	0	1	--
Carbon	Aromatic	Side Chain	TRP	CG,CD1, CD2,	0	1	--
Carbon	Aromatic	Side Chain	TRP	CE2, CE3, CZ2, CZ3, CH2	0	1	--
Sulfur	Sulfhydril	Side Chain	CYS	SG	0	-1	H Bond
Sulfur	Thioether	Side Chain	MET	SD	0	0	--
Oxygen	Sulfate	HET Group	SO4	O1, O2, O3, O4	-ve	-1	H Bond
Oxygen	Phosphate	HET Group	2HP	O1, O2, O3, O4	-ve	-1	H Bond
Oxygen	Water	HET Group	HOH	O	0	-1	H Bond
Calcium	Ion	HET Group	CA	CA	+ve	-1	H Bond
Magnesium	Ion	HET Group	MG	MG	+ve	-1	H Bond
Zinc	Ion	HET Group	ZN	ZN	+ve	-1	H Bond

Set A = ALA, SER, THR, CYS, ASP, ASN, GLU, GLN, ARG, LYS, PRO

Lacking differential rule ground truth, we can not use this method for real world data. Uplift curves are often used to address this problem [111]. Using 5-fold cross-validation, we use the learned theory rules as attributes to a Bayes Net TAN classifier [47] 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 $\{p, L_1 - L_2\}$.

3.6 Augmenting a Bayes Net with Rules

Bayesian Belief Networks (Bayes Nets for short) are informatics tools used for predicting the probability (risk) of an outcome based on observed variables. Bayes Nets predict the probability of an outcome using a graphical structure encoding variables (nodes), conditional dependence relationships (arcs) and probabilities quantified in conditional probability tables associated with each node [79].

Given feature vectors composed of discrete variables, Bayes Nets can be learned directly from data. Using various heuristic search techniques, the objective is to infer a network that best represents the training data probability distribution, as measured by likelihood, BIC, or other measures of fit to data. After the Bayes Net structure is determined, conditional probability tables are computed using standard occurrence counting techniques [79].

We use the Tree Augmented Naive Bayes (TAN) algorithm [47], as implemented in Weka [55]. TAN starts with a Naive Bayes structure: the class variable has no parents, and is itself the sole parent of each attribute. TAN then adds arcs between variables to approximate the interactions between attributes. It uses a tree structure to ensure that each attribute has at most one other attribute augmenting edge pointing to it.

Rules can be incorporated into a Bayes Net as additional variables in the original feature vector data. Each rule can be seen as a binary variable: a given example is either covered or not covered by that rule. We then learn a Bayes Net over the rule-augmented data.

Chapter 4

The Model Filtering Approach

This chapter *introduces our automated Model Filtering (MF) approach.* Applying this method to breast cancer resulted in *the first instance of differential predictive rules discovery.* This chapter is based on a paper presented at *ACM International Health Informatics (IHI-10)* [89], and on another submitted to *PLoS ONE* journal [8].

4.1 Problem Motivation

Breast cancer is the most common type of cancer among women. An estimated 1.3 million new cases of invasive breast cancer were expected to occur among women in 2007 [49]. Statistical data shows that a woman in the US has a 1/8 lifetime risk of developing breast cancer [4].

There are two basic stages of breast cancer. If cancer cells are confined within the ducts and lobules where they developed and have not spread, the stage is *in situ.* If cancer cells have broken through their originating ductal or lobular structures to invade the surrounding tissue, the stage is *invasive.* In situ cancers represent 25% of breast tumors [3].

Since nearly all in situ cases can be cured [3], current practice is to treat in situ occurrences in order to avoid progression into invasive tumors [4]. 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 [2].

Researchers have successfully used pre-biopsy mammography features to build breast-cancer classifiers capable of discriminating between invasive and in situ cancers [66, 85, 128]. Mammography, or x-ray of the breast, is the main tool used for early detection of breast cancer. A routine asymptomatic mammography exam is called a *screening mammogram*, while a more detailed exam following symptoms or a higher risk is called a *diagnostic mammogram*.

Using patient characteristics and mammography findings to uncover invasive and in situ differential predictive rules may help decrease the number of biopsies which may confer minimal benefit in elderly women; and target interventions to younger women, who would benefit the most from early diagnosis.

4.2 Age Matters

To accentuate age-based differences, we limit our age-based analysis to the mammography younger and older cohorts (Table 3.1). If age based differences exist, they are most likely explained by steady and gradual changes rather than an abrupt shift at any single age. In fact, early work showed that the assignment of mammography exams into specific age cohorts with a certain cut point (usually at age 50) may not be desirable unless outcomes abruptly change at this cut point [68]. Changes due to menopause do not appear as sharp changes at any specific age when averaged over a population of women. Removing the middle-age group helps impose a more marked distinction between older and younger age groups making potential observed differences clearer.

To identify differential prediction in the context of age, we fit a Multivariable Logistic Regression model for the older and younger cohorts using the structured and extracted features of Table 3.2, and compare the resulting curves. We use stepwise logistic regression [64], which minimizes the Akaike Information Criterion (AIC) goodness of fit measure to eliminate predictors that do not substantially improve the model fit. The pairwise correlations among the predictors are small and the results from the stepwise fits are stable. We do not include interaction terms because sparse information in the two way tables between predictors made these terms difficult to estimate and interpret. We evaluate the performance of our models using leave-one-out cross validation.

The resulting older women regression model includes eight variables, six of which are statistically significant at the 95% level in predicting invasive cancer versus in situ (Appendix Table B.1). Specifically, presence of a palpable lump ($p = 0.013$), family history of breast cancer ($p = 0.043$), principal abnormal finding ($p < 0.001$), calcification distribution ($p = 0.008$), mass margins ($p < 0.001$), and mass shape ($p = 0.033$) are statistically significant. Prior surgery ($p = 0.132$) and focal asymmetric density ($p = 0.077$) were included but were not statistically significant.

The resulting younger women regression model includes five variables, three of which were statistically significant at the 95% level (Appendix Table B.2). Presence of a palpable lump ($p < 0.001$), principal abnormal finding ($p < 0.001$), and mass size ($p = 0.047$) were significant. Architectural distortion ($p = 0.063$) and mass shape ($p = 0.090$) were included but were not statistically significant.

We compare the performance of both models in predicting cancer stage (invasive versus in situ) using the Area Under the ROC Curve (AUC-ROC) value (Figure 4.1). The model for older women achieved an AUC-ROC of 0.848 whereas the model for younger women had an AUC-ROC of 0.778, a significant difference ($p = 0.049$).

These results show that the invasive versus in situ classification problem exhibits age-based differential prediction. The predictive ability of our logistic regression models in differentiating between invasive and in situ does depend on age. Even though both models

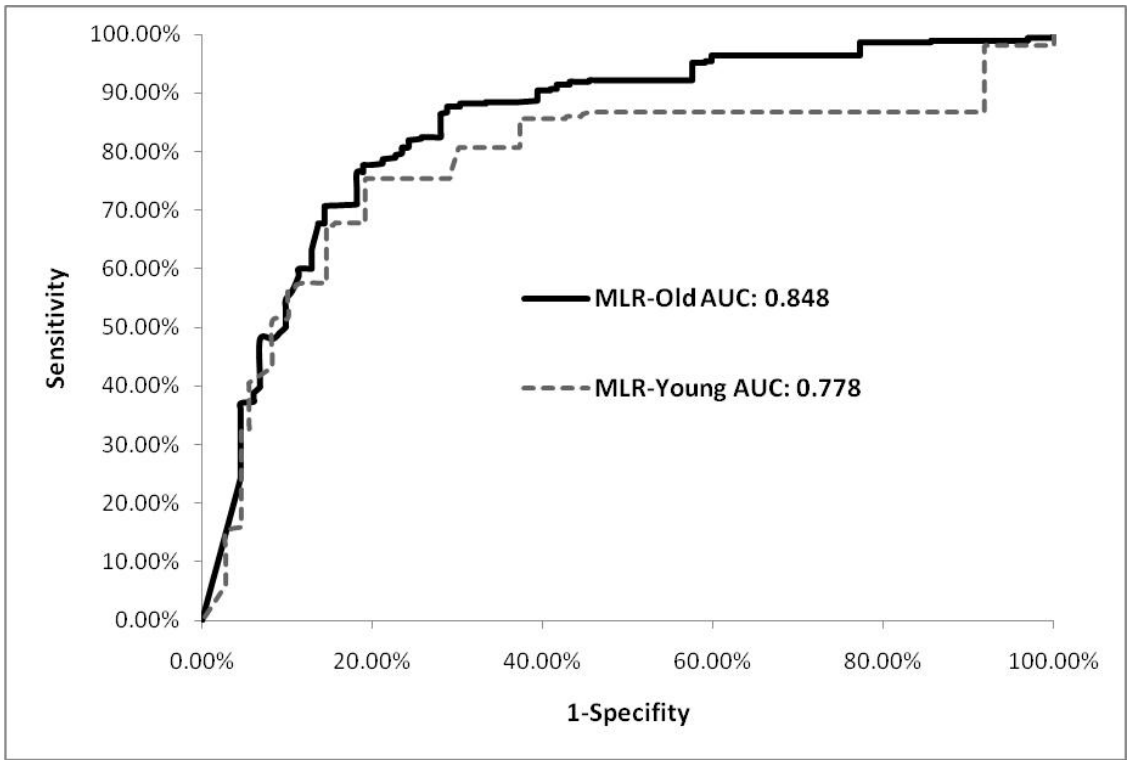


Figure 4.1: ROC curves and area under the curve (AUC) for old and young patients Multivariable Logistic Regression models

found the presence of a palpable lump and the principal abnormal finding to be significant predictors, each regression incorporated different additional significant predictors. Family history, calcification distribution, mass shape, and mass margins for older, and mass size for younger. In addition, our older women model significantly outperforms our younger women model, when each is applied on its data subset.

4.3 Model Filtering Method

Having established differential prediction and the fact that, based on age groups, different mammographic features can be used to classify cancer as invasive or in situ, we now shift our attention to extracting differential predictive rules.

We start by constructing an ILP model over the target stratum. The search is guided by a clause evaluation function that maximizes some statistic (e.g. coverage, compression, entropy) discriminating between invasive and in situ over the given target stratum. The model outputs a high-performance stratum-specific theory. By construction, the theory rules perform well on their stratum, according to the 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 approach, the greater the performance difference, the more differentially predictive a rule is.

As an example, the flowchart of Figure 4.2 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 learner generates a theory composed of rules which, by construction and taken together, explain the training data. The generated rules are expected to have a good performance over the older stratum. We then test each rule on the younger stratum, and keep rules that perform poorly.

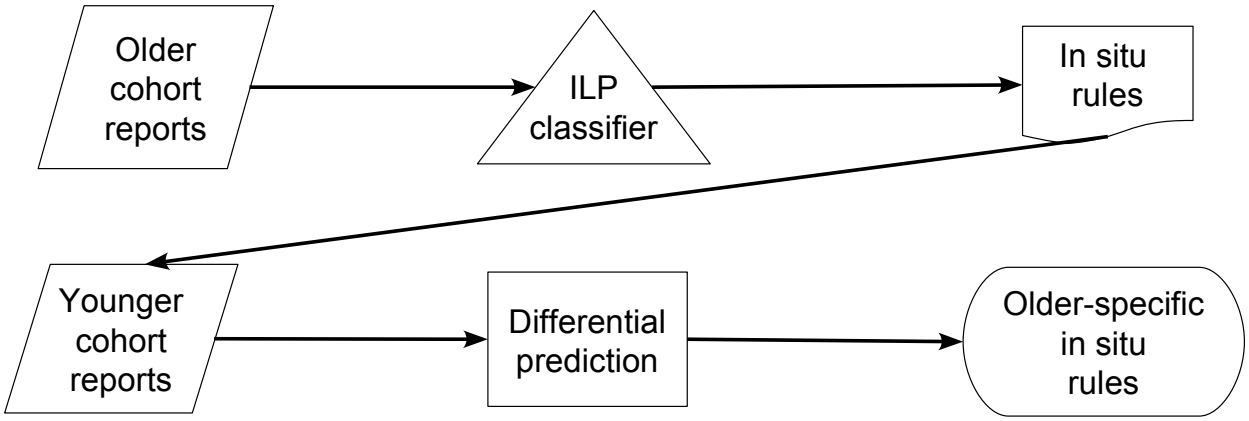


Figure 4.2: Model Filtering approach to identify older-specific in situ rules

The differential prediction rule filtering component can be based on a statistical significance test, on a threshold, on cross-validation or on a tuning set. For example, we can set aside older and younger tuning sets, and select rules whose precision is significantly worse on the younger tuning set when compared to the older.

4.4 Experiments and Results

To test our approach, we apply it to each age (older, younger) and cancer stage (in situ, invasive) combination, using the structured, extracted (Table 3.2) and extensional (Table 3.3) features. We opt for a tuning set model filter, and divide each cohort in half to form a training and a tuning set. We make sure all records pertaining to the same patient end up in the same subset. We perform our experiments using Aleph [121] running within the Yap Prolog compiler [115].

To avoid reporting rules with low coverage, or with bad predictive precision, we select rules whose recall on the same-age tuning subset is greater than or equal to 10%, and precision on the same-age tuning subset is greater than or equal to 60%. We filter and select rules whose precision is significantly better, at the 95% confidence level, on one tuning subset compared to the other. We assume a uniform prior and use a probabilistic interpretation of precision in order to compare precision measurements on different datasets [53].

We present each logical rule's English translation. We divide them by the age cohort and cancer stage categories to which they refer. In cases where no rule meets our selection criteria for a certain category, we report sub-optimal rules for completeness as well as comparison purposes. We group rules by predicate similarity and provide their clinical summary. We include for each rule its invasive and in situ coverage, as well as its precision p and recall r, on both its corresponding younger and older tuning subsets. We also include the results over the middle cohort for comparison purposes. The middle cohort experiments were not used in rule generation or selection.

4.4.1 Rules Predicting Invasive in Older Cohort

The following invasive-predicting rules have a significantly better precision, at the 95% confidence level, on the older cohort when compared to the younger. A cancerous diagnostic mammogram A is invasive if:

1. **The mammogram has a palpable lump in this-side breast.**
   (younger: 86 invasive, 13 in situ, $p = 87\%$, $r = 65\%$)
   (middle: 99 invasive, 15 in situ, $p = 87\%$, $r = 50\%$)
   (older: 85 invasive, 5 in situ, $p = 94\%$, $r = 42\%$)

2. **The mammogram's indication for exam is "breast problem palpable lump".**
   (younger: 82 invasive, 13 in situ, $p = 86\%$, $r = 62\%$)
   (middle: 78 invasive, 15 in situ, $p = 84\%$, $r = 39\%$)
   (older: 71 invasive, 4 in situ, $p = 95\%$, $r = 35\%$)

3. **The mammogram's indication for exam is "breast problem palpable lump",**
   its other side BI-RADS score is less than 3,
   and its mass margin is not reported.
   (younger: 54 invasive, 8 in situ, $p = 87\%$, $r = 41\%$)
   (middle: 42 invasive, 6 in situ, $p = 88\%$, $r = 21\%$)
   (older: 39 invasive, 1 in situ, $p = 98\%$, $r = 19\%$)

These three rules show that the presence of a palpable lump leads to a more precise prediction of invasive cancer as compared to in situ in older women. Having a palpable lump in younger women does not differentiate as well between invasive and in situ.

4. **The mammogram has an old-biopsy that was invasive.**
   (younger: 24 invasive, 4 in situ, $p = 86\%$, $r = 18\%$)
   (middle: 82 invasive, 1 in situ, $p = 99\%$, $r = 41\%$)
   (older: 101 invasive, 3 in situ, $p = 97\%$, $r = 50\%$)

5. **The mammogram has an old-biopsy that was invasive,**
   and the biopsy happened within the same age group.
   (I.e. an older women had the prior biopsy when she was above 65 years old)
   (younger: 24 invasive, 4 in situ, $p = 86\%$, $r = 18\%$)

(middle: 81 invasive, 0 in situ, $p = 100\%$, $r = 41\%$)
(older: 89 invasive, 0 in situ, $p = 100\%$, $r = 44\%$)

In the setting of recurrence, older women may be more likely to have invasive rather than in situ cancer. In other words, the fact that a woman is having a recurrence is a better predictor of invasiveness in older women than it is in younger women.

4.4.2 Rules Predicting In Situ in Older Cohort

Only one in situ-predicting rule has a significantly better precision, at the 95% confidence level, on the older cohort when compared to the younger. Its recall value is 4.55%, well below our cutoff value of 10%, and thus is a sub-optimal rule that we report for completeness. A cancerous diagnostic mammogram A is in situ if:

1. **The mammogram's indication for exam is "breast problem other",**
   there is no prior surgery,
   and its mass size is not reported.
   (younger: 2 in situ, 7 invasive, $p = 22\%$, $r = 4\%$)
   (middle: 11 in situ, 9 invasive, $p = 55\%$, $r = 13\%$)
   (older: 3 in situ, 1 invasive, $p = 75\%$, $r = 5\%$)

This rule's coverage is very low and doesn't allow for an adequate clinical interpretation.

4.4.3 Rules Predicting Invasive in Younger Cohort

No invasive-predicting rule has a significantly better precision, at the 95% confidence level, on the younger cohort when compared to the older. The best discriminating rule is only significant at the 87% confidence level, and is thus a sub-optimal rule. A cancerous diagnostic mammogram A is invasive if:

1. **The mammogram has a palpable-lump in this-side breast,**
   its breast density is class 2,

and its calcification distribution is not reported.
(younger: 15 invasive, 1 in situ, $p = 94\%$, $r = 11\%$)
(middle: 31 invasive, 0 in situ, $p = 100\%$, $r = 16\%$)
(older: 23 invasive, 6 in situ, $p = 79\%$, $r = 12\%$)

Low breast density usually allows for easier mass detection on the mammogram. However, when there is a palpable finding, the detection task facilitated by low breast density ceases to be important.

4.4.4 Rules Predicting In Situ in Younger Cohort

The following in situ-predicting rule has a significantly better precision, at the 95% confidence level, on the younger cohort when compared to the older. A cancerous diagnostic mammogram A is in situ if:

1. **The mammogram has a personal history of cancer in this-side breast,**
   this-side breast has a prior surgery,
   and its combined BI-RADS increased by at least 2 points compared to a previous study.
   (younger: 6 in situ, 3 invasive, $p = 67\%$, $r = 11\%$)
   (middle: 4 in situ, 9 invasive, $p = 31\%$, $r = 5\%$)
   (older: 1 in situ, 11 invasive, $p = 8\%$, $r = 2\%$)

This rule suggests that if a patient has a recurrence, this is a better predictor of in situ in younger women. This rule complements rules $4 - 5$ in Section 4.4.1.

4.5 Differential Rules Discussion

Our MF differential prediction approach provides a number of interesting rules, some of which are previously unreported and are worthy of further investigation.

4.5.1 Predicting Invasive in Older Cohort

Starting with rules predicting invasive in older women, we notice that the first three rules involve palpable lump, the first two rules having it as a sole predicate. We further check recall values, and find that the three rule's recall is significantly better for the younger cohort. This means that there is a significantly higher percentage of younger women diagnosed with palpable lumps; but the presence of a palpable lump is a significantly more precise indicator of invasiveness in older women.

Typically women under the age of 40 are not included in a breast-screening program. Because younger women with breast cancer rarely undergo mammography before diagnosis, they often present a palpable lump detected through self-examination or by assessment by their general practitioner [48]. As opposed to a screening mammogram detection, which is often the case with older women. This explains higher palpable lump recalls associated with the younger cohort.

The palpable lump rules' higher precision associated with the older cohort is more interesting. Here is a possible explanation. Studies have shown that breast cancer in younger women is pathophysiologically more aggressive and has a poorer prognosis [38, 48]. Younger women tend to have higher proportions of poorly differentiated, rapidly proliferating tumors that tend to be larger and to involve regional lymph nodes [1]. Due to their larger size, the tumors are more likely to be palpable, increasing the palpability likelihood of an in situ tumor in younger women. Which may explain the palpable lump rules' better precision over the older cohort, where the mass grows at a slower pace, and once it is big enough to be palpable, it is almost certainly invasive. These rules merit further investigation, with a possible factoring of histological grade and date of last screening mammogram.

Rules 4 and 5 predict an invasive tumor based on a prior biopsied invasive tumor. Both rules also exhibit a significantly better recall in the older cohort. This reflects the higher risk of proliferation and recurrence of invasive tumors [74] which, combined with a longer life-span for the recurrence to manifest itself, is more common in older women.

4.5.2 Predicting In Situ in Older Cohort

The only reported rule is based on a very small number of older examples and doesn't meet the 10% recall cut-off. It specifies "other" as the clinical indication for the exam, a miscellaneous and not very informative category. In addition, rules reporting the absence of features are difficult to clinically interpret. Unfolding in situ-predicting rules with a significantly better performance in older women requires further studies.

4.5.3 Predicting Invasive in Younger Cohort

Although the reported rule is only significant at the 87% confidence level, and no conclusions should be drawn based on it, it sheds some light on the previously discussed palpable lump issue.

The rule requires a palpable lump in this-side breast, together with a breast density of class 2, scattered fibroglandular tissue. This is a relatively low breast density for younger women, since it is well established that younger women tend to have denser breasts than older women [63, 135]. Mammogram sensitivity significantly increases with declining breast density [78], since a low breast density allows for easier mass detection on the mammogram. However, when there is a palpable finding, the detection task facilitated by low breast density ceases to be important. While the discriminating ability of low breast density may explain the relative increase in invasive detection precision in younger women in this rule, the inclusion of a palpable lump predicate adds some doubts to the clinical explanation of this rule.

4.5.4 Predicting In Situ in Younger Cohort

The rule predicting in situ in the younger cohort requires both a prior surgery and a personal cancer history to be present in the same breast. Combined with a BI-RADS increase, it favors in situ in younger and invasive in older. This rule complements rules $4-5$ in Section 4.4.1, suggesting that a recurrence is a better predictor of in situ in younger women.

This rule covers more invasive than in situ cases when tested on the older subset. It thus provides opposite predictions across the age divide. In addition, it is the only rule that links the current mammogram to older ones. This rule takes full advantage of ILP's relational capabilities, and allows previous mammograms features to influence the current mammogram classification.

Opposite predictions across age-strata, and linking to previous mammograms, this previously unreported rule offers a clear-cut age-specific personalized prediction and merits further clinical investigation.

4.6 Middle Cohort Comparison

To accentuate age-based differences, we limited our differential rules generation to the younger and older cohorts. In this section, we investigate the performance of the resulting rules on the middle cohort. Table 4.1 compares the middle cohort performance of each differential rule (see Section 4.4) to its performance on the younger and older cohorts. We apply the same statistical test used to select our age-specific rules.

Suppose a rule has a middle cohort performance that is significantly different from one non-middle cohort, say older, and is not significantly different from the other, younger in this case. Then the middle cohort is more similar to the non-significant (i.e. younger) cohort in the scope of the concerned rule. On the other hand, suppose a rule has a middle cohort performance that is not significantly different from both non-middle cohorts. Then, in the scope of this rule, the middle cohort shares similarities with, and its features lie in between, the two other cohorts. For our age-specific rules, the middle cohort behaves indeed as a "middle" cohort. For some rules it displays similarities to either the younger or the older cohorts, while in others it is situated in the middle.

4.7 Model Filtering Approach Discussion

Our results show that the MF approach works. It is able to learn meaningful differential predictive rules. The resulting rules are age-specific, for which the discarded middle cohort

Table 4.1: Middle Cohort Precision Comparisons

	Comparing Middle Cohort with:	
Rule (see Section 4.4)	Older Cohort (p-value)	Younger Cohort (p-value)
Invasive Older Prediction		
Rule 1	0.04*	0.50
Rule 2	0.01*	0.32
Rule 3	0.05	0.49
Rule 4	0.26	0.00*
Rule 5	0.48	0.00*
In Situ Older Prediction		
Rule 1	0.27	0.06
Invasive Younger Prediction		
Rule 1	0.00*	0.12
In Situ Younger Prediction		
Rule 1	0.10	0.06

* Statistically significant at the 95% confidence level.

behaves indeed as a "middle" cohort. Nevertheless, the coverage of some of the rules is small, and filtering rules solely on their differential precision may not be adequate. In addition, we did not uncover any significant nor meaningful older in situ differential rule, which is our main motivation.

In order to infer meaningful and significant older in situ rules, we may need to use a different filtering function, one that also incorporates size and recall. In addition, we can increase the training set size by using a statistical test differential filter instead of a tuning set. Finally, we may need a different approach more tuned to the differential prediction search space. We explore these alternatives in the next chapter.

Chapter 5

Differential Prediction Search Approach

The Model Filtering (MF) approach is a generate-then-test method. Target stratum rules are generated by training solely on the target stratum subset, and then filtered by tuning on the other stratum. A more rigorous approach is to use test-incorporation. This chapter *introduces the Differential Prediction Search (DPS) approach*, which builds a differential prediction classifier by altering the ILP search space. We *test and compare our methods on synthetic data as well as on the mammography data*. We establish that *for large and noisy data, which is what most real world applications are, DPS is more appropriate*. Finally, we *augment a Bayes Net with differential rules* for risk prediction, forming a Logical Differential Prediction Bayes Net (LDP-BN), and observe a significant performance increase. This chapter is based on a paper presented at *ECML-12* [90], and on another accepted at *AMIA-12* [92].

5.1 Differential Prediction Search Method

Our third method, *differential prediction search* (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, rules 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 5.1.1 (**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.

A very simple example is the difference of the classification score of a rule over both strata. Let $S(R|D_i)$ be the performance score for R over the subset D_i. We have:

$$Q(R|D_t, D_o) = S(R|D_t) - S(R|D_o). \tag{5.1}$$

Figure 5.1 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.

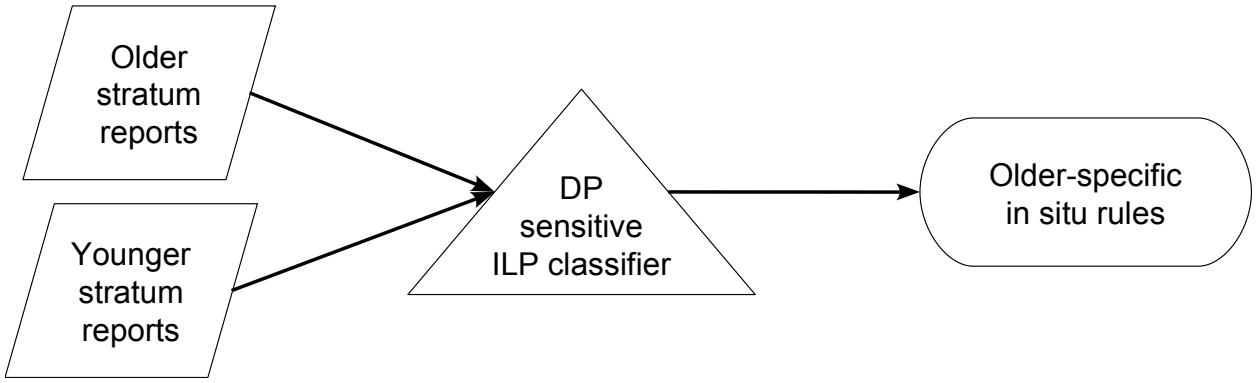


Figure 5.1: Differential prediction search approach to identify older-specific in situ rules

5.2 Scoring Functions

The approaches we propose can be applied to any ILP algorithm, and can be used with any scoring function S. In this work, we use the 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. Given a rule R covering $P(R)$ positives and $N(R)$ negatives over data D, with $Prior$ being the fraction of positive examples in the data D, m-estimate is computed as:

$$mEstimate(R|D) = (P(R) + m \times Prior) \div (P(R) + N(R) + m)). \tag{5.2}$$

An important concern in real-life situations is population size [104]. 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 multiplying the m-estimate score by the rule positive cover, as shown below for each approach.

We implement our differential predictive rule learning methods using Aleph [121]. 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 [115].

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. We do not hold out tuning sets.

5.2.1 Baseline Score

We first establish a standard ILP baseline. We 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:

$$S_{BASE}(R|D_t, D_o) = poscover(R|D_t \cup D_o) \times mEstimate(R|D_t \cup D_o). \quad (5.3)$$

5.2.2 Model Filtering Score

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:

$$S_{MF}(R_t|D_t, D_o) = S_{BASE}(R_t|D_t) - S_{BASE}(R_t|D_o). \quad (5.4)$$

5.2.3 Differential Prediction Search Score

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

$$Q_{DPS}(R|D_t, D_o) = poscover(R|D_t) \times (mEstimate(R|D_t) - mEstimate(R|D_o)). \quad (5.5)$$

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 [111]. Postulate 2 is trivially satisfied: if the condition is independent from treatment then 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.

5.2.4 Instance Relabeling Score

Finally, we consider the Instance Relabeling method (Section 2.3.3). This method is specific to the coverage rule-scoring evaluation function; we do not use m-estimate and compensate for population size as we did for the other methods.

Let P_t and N_t be the positive and negative instances of the target stratum (older in our case), and P_o and N_o be the positive and negative instances of the other stratum (younger in our case). A given rule R scores:

$$S_{RLBL}(R|D_t, D_o) = cover(R|P_t \cup N_o) - cover(R|P_o \cup N_t). \tag{5.6}$$

5.3 Michalski-Trains Results

Before going to our mammography target application, we use Section 3.3 Michalski-Trains 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 data has 30 simulations for each scenario, noise level, size and method combination. Table 5.1 reports the AUC-PR mean and standard deviation for each method and experimental block. The Instance Relabeling method runs took too long, and retrieved few true positive rules, if any. We discard this approach after a few simulations. See Section 2.3.3 for a discussion of this method's weakness.

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). See Table 5.2 for detailed comparisons.

Table 5.1: 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

Table 5.2: p-value of pairwise Hommel adjusted paired two-tailed Wilcoxon tests. Significant results are in bold.

	Clean				Noisy			
	100		1000		100		1000	
	BASE	DPS	BASE	DPS	BASE	DPS	BASE	DPS
One target rule scenario								
DPS	0.33	-	0.83	-	1	-	**0.03**	-
MF	0.33	**0.02**	0.83	0.83	1	1	0.06	0.44
Multiple target rules scenario								
DPS	**0.03**	-	0.65	-	0.30	-	**0.04**	-
MF	0.14	**0.00**	0.14	0.21	0.21	**0.00**	0.56	**0.05**
Both scenarios combined								
DPS	**0.03**	-	0.79	-	0.41	-	**0.00**	-
MF	**0.05**	**0.00**	0.26	0.20	0.13	**0.02**	**0.04**	**0.04**

5.3.1 Michalski-Trains 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 of the DPS losses 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 5.1). 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 the DPS increases being significant. In these experiments, DPS increases an average of 32 percentage points, compared to 12 for BASE and 11 for MF (Table 5.1). Although more experiments are necessary to establish a performance-size curve, this evidence suggests that BASE and MF increase rate may be stalling at size 1000, while DPS performance is likely to keep improving.

Although no clear pattern emerges from comparing different methods on one-target versus 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 5.1). 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* (Table 5.2). 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.

5.4 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 described in Section 3.2. We use the same experimental setting as for the synthetic data, but set $nodes = 200,000$ since the number of predicates is much larger.

5.4.1 Breast Cancer Diagnosis Results

The BASE method does not return any rules, which highlights the difficulty of this task. Lacking ground truth, we use uplift curves (see Section 3.5) to compare MF and DPS (Figure 5.2). 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.

As the data sets are skewed, we present the precision and recall curves for the classifiers in Figure 5.3. In both cases, the differential rules theory performs better on the older cohort, as it should. This is the case almost across the whole PR space, even though the older cohort has a weaker negative bias. The exception is for very high recalls, where the difference in bias dominates. We also note a larger gain for the DPS method.

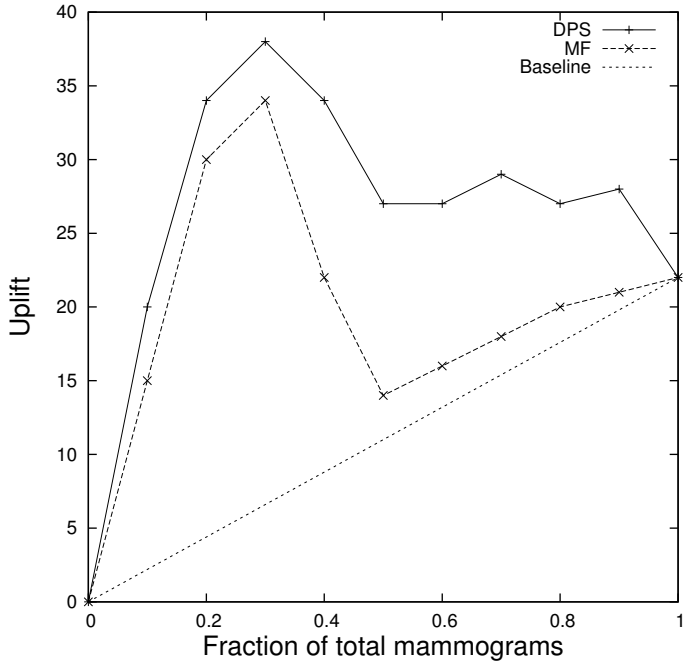
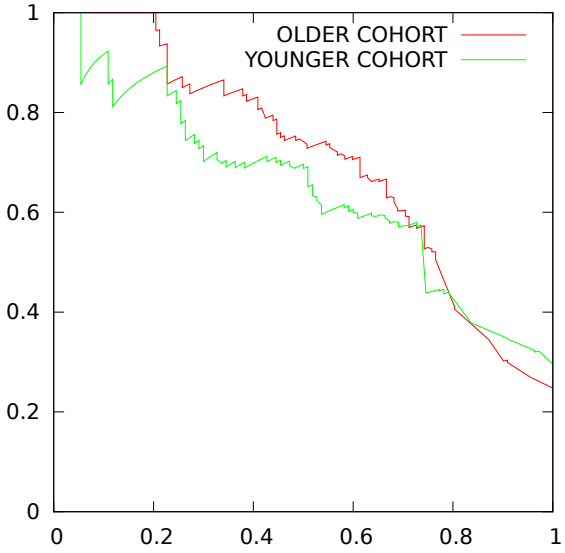


Figure 5.2: Uplift curve for breast cancer stage (see Section 3.5)

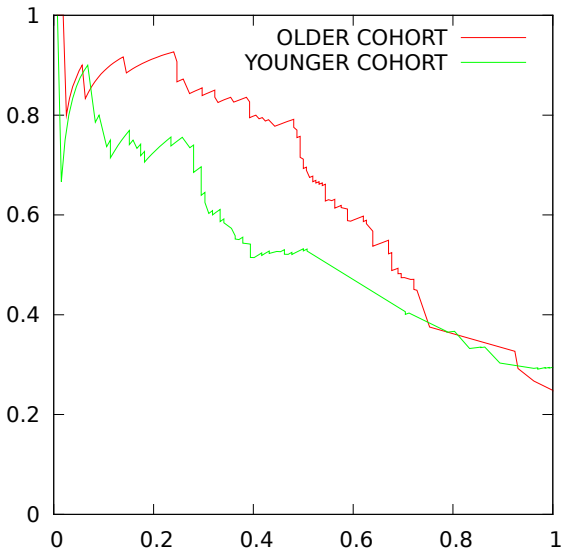
Uplift curves provide good insight into how the different models differ in terms of lift. To obtain more insight into theory quality and how the theories differ between folds and between cohorts, we compare the per-fold AUC-PR of the differential theories across both cohorts (Table 5.3). In this case the differences are less clear for MF, but DPS consistently shows an increase from younger to older. We also observe that DPS tends to generate theories that perform better on the older cohort and worse on the younger cohort. We performed a t-test on the 5 folds, and the difference between cohorts is significant at the 99% confidence level for DPS, and not significant for MF at the 95% confidence level .

5.4.2 Breast Cancer Differential Rules

MF returns 4 differential predictive rules that have a significantly better precision and recall [53] 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:



(a) MF PR Curve



(b) DPS PR Curve

Figure 5.3: Pooled Precision and Recall curves for the MF and DPS methods on the two age cohorts

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.

Table 5.3: AUC-PR difference between the two cohorts per fold

	MF			DPS		
	older	younger	Δ	older	younger	Δ
1	0.74	0.64	0.10	0.76	0.54	0.22
2	0.86	0.82	0.04	0.84	0.64	0.20
3	0.62	0.84	-0.22	0.89	0.80	0.09
4	0.71	0.68	0.03	0.80	0.69	0.11
5	0.63	0.64	-0.01	0.78	0.51	0.27

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 [97]. Younger women tend to have more rapidly proliferating cancers that develop into a palpable mass [48], 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 [48]. 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 [130].

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 [5]. 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 [48]. Thus older women tend to have more screening exams because of regular visits after age 40.

Finally we note the relevance of 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 [63]. 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 [78], allowing for easier micro-calcification detection.

5.5 Logical Differential Prediction Bayes Net

In order to advocate for watchful waiting rather than biopsy in women > 65, risk prediction of benign, in situ, and invasive disease based on mammographic features must be accurate. The literature confirms that the mammographic appearance as described by the radiologist can predict the histology of breast cancer [128, 126]. Fortunately, mammography performs superiorly in older women [108]. In fact, Bayes Net models built using BI-RADS mammography features can accurately determine breast disease in a general population [18, 19].

5.5.1 Augmenting a Bayes Net with Differential Rules

Nevertheless, to personalize and optimize breast cancer diagnosis specifically to aging women, we need multirelational algorithms that can address the reality of disease heterogeneity (in our case, based on age), while learning predictive variables for risk prediction in the target population. We can achieve this by combining differential predictive rules with a Bayes Net (see Section 3.6), thus forming a Logical Differential Prediction Bayes Net (LDP-BN). LDP-BN includes: 1) leveraging multi-relational data to discover predictive rules via Inductive Logic Programming (ILP), 2) addressing breast cancer heterogeneity by performing differential prediction over age, and 3) incorporating these predictive logical rules, tailored to women > 65, into a Bayes Net for classification/risk prediction.

Once we generate ILP differential rules, we incorporate them as additional variables in the original feature vector data. We then learn a Bayes Net over the rule-augmented data. This resulting Bayes Net is a Logical Differential Prediction Bayes Net.

5.5.2 LDP-BN Results

We compare four different Bayes Nets over the older-stratum mammography data. The first is a baseline Bayes Net built without the added differential prediction rules. The second is an LDP-BN that uses older-specific rules learned using the MF method. The third is an LDP-BN that uses older-specific rules learned using the DPS method. The fourth is a Bayes Net augmented with non-differential rules generated by Aleph over the older cohort.

To train and test our Bayes Nets, we use conventional stratified 10-fold cross validation. We construct the ROC curves with the final curve being the result of vertically averaging the 10 curves from the 10 folds. For rule generation, we use the same scoring functions and experimental setting as Sections 5.2 and 3.5 above. We select rules whose older-stratum precision and recall results are each no worse than younger's, with one of them being statistically significantly better at the 95% confidence level [53]. We mine both invasive and in situ older-specific differential predictive rules and incorporate all of them into the Bayes Net for older-specific invasive/in situ prediction.

Table 5.4 shows the Area Under the ROC Curve (AUC-ROC) for each of the 10 folds. Figure 5.4 depicts the final ROC curves. We note that the DPS augmented Bayes Net is constantly outperforming the MF augmented one, which in turn constantly outperforms the Baseline Bayes Net. In fact, the differences are statistically significant at the 99% confidence level. A paired two-tailed t-test gives p-value < 0.0001 for MF compared to Baseline, and p-value $= 0.0055$ for DPS compared to MF.

The non-differential Aleph-augmented Bayes Net is constantly outperforming MF, and the difference is statistically significant (p-value $= 0.0041$). Although the Aleph average is better than its DPS counterpart, their two ROC curves cross, and the difference is not significant (p-value $= 0.7388$). DPS performs better at a lower false positive rate, and Aleph

Table 5.4: Area under the ROC curve results for the baseline, MF, DPS and Aleph augmented Bayes Nets over the 10 folds

Fold	Baseline BN	MF LDP-BN	DPS LDP-BN	Aleph BN
1	0.8067	0.8846	0.9587	0.8942
2	0.8208	0.8996	0.9062	0.9644
3	0.8714	0.8973	0.9482	0.9357
4	0.8438	0.8714	0.9080	0.9187
5	0.7990	0.8615	0.9000	0.9096
6	0.8769	0.9346	0.9615	0.9452
7	0.7183	0.8163	0.8337	0.8481
8	0.9154	0.9654	0.9442	0.9654
9	0.8490	0.9317	0.9558	0.9279
10	0.8154	0.8779	0.9365	0.9067
Average	0.8304	0.8911	0.9197	0.9212

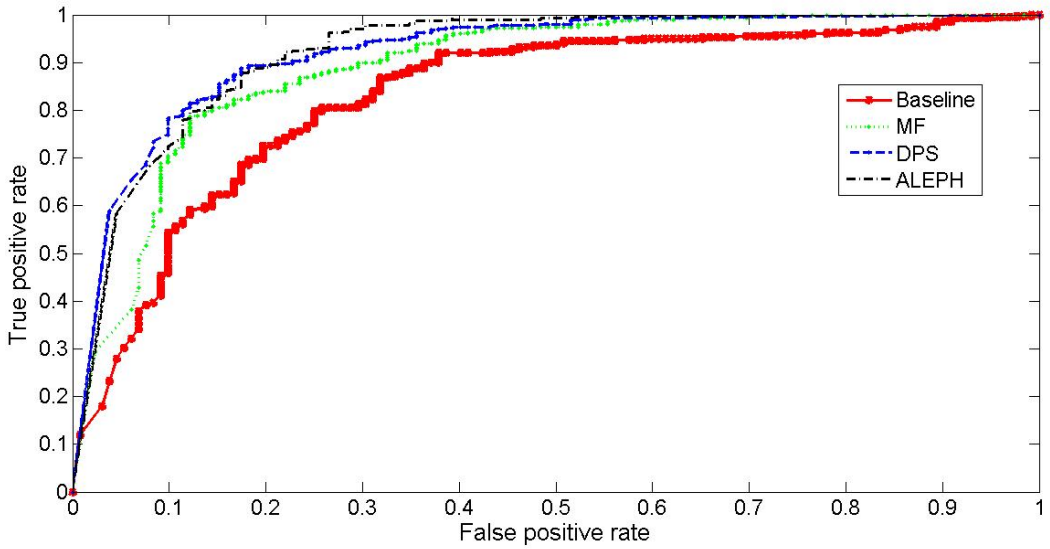


Figure 5.4: Final ROC curves for the baseline, MF, DPS and Aleph augmented Bayes Nets

at a higher false positive rate. The two curves intersect at a true positive rate (recall, sensitivity) of 0.9 and at a false positive rate (1-specificity) of 0.2. Even though Aleph-augmented and DPS-augmented Bayes Nets have similar performances, DPS LDP-BN has the advantage of offering differential prediction insight into the underlying domain.

It is important to note that LDP-BN rules are learned for their differential predictive potential, separately from the Bayes Net. The differential rules identification and the Bayesian Network construction are not integrated into a global optimization framework, as in the SAYU system [32]. It may be possible to further improve the Logical Differential Prediction Bayes Net by doing so in the future.

Chapter 6

The Expert Driven Approach

So far we introduced two automated differential predictive rule learning methods. A more basic approach is to rely on an expert to infer differential predictive rules by comparing models built on the different strata. In this chapter, we *introduce the Expert Driven (ED) approach for differential prediction.* Applying the Expert Driven approach to extract older-specific in situ breast cancer stage rules resulted in too many rules to be practical. For illustration, we apply it to infer differences between specific glucose and general hexose binding, showing that it works on an important biological problem. In doing so, we create the *first glucose-binding classifier*, and perform an ILP-based *data-driven empirical validation of biochemical hexose-binding knowledge.* This chapter is based on a paper published in the journal *Proteins: Structure, Function and Bioinformatics* [86], and on another presented at the *ILP-09* conference [87].

6.1 Biological Background

We start by an overview of hexoses and their binding properties.

6.1.1 Problem Significance

Hexoses are 6-carbon sugar molecules that play a key role in several important biochemical pathways, including cellular energy release, signaling, carbohydrate synthesis, and the regulation of gene expression [120]. Proteins that bind these sugars are implicated in several human diseases, including diabetes, various metabolic disorders, and Huntington disease.

The biochemical and molecular pathways for these disease mechanisms have not all been elucidated and much work remains to be done.

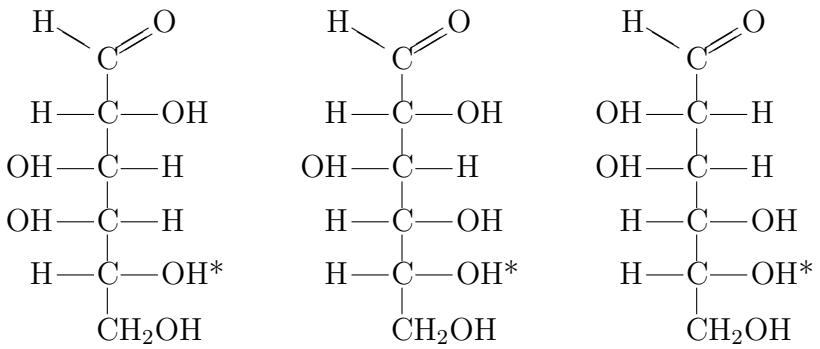
In parallel, genome sequencing of a wide range of species has yielded sequence knowledge of a large number of proteins whose biochemical functions are still unknown. The three-dimensional structures of many of these proteins were elucidated. Some of these proteins have been shown to be members of certain pathways, but they lack sufficient sequence or structural similarity to any other protein with a known function.

The functional annotation of these "unknown" proteins is of paramount importance. One approach to tackle this problem is to predict what these proteins may bind to. Prediction of glucose-specific binding sites will significantly improve our understanding of protein structure-function relationships and enable us to assign possible functions to some of the many genomic proteins whose function remains unknown [98]. This, in turn, will allow us to better understand disease mechanisms that may involve some of these proteins and be better placed for either diagnosis or treatment of these diseases.

6.1.2 Hexoses

Galactose, glucose and mannose (Figure 6.1) are, in this order, the most commonly found hexoses in nature [44]. All hexoses have two chemical groups that can react together, the carbonyl group and the hydroxyl group situated on carbon number 5 (see Figures 6.1 and 6.2). The reaction between these two groups folds the molecule on itself as shown in Figure 6.3. This intra-molecular cyclization reaction forms a *pyranose ring* from five carbons and one oxygen atoms [20]. The cyclized hexose can adopt either of two configurations, α or β, according to whether the hydroxyl group $-\mathrm{OH}^*$ is located below or above the pyranose ring.

Hexoses can readily shift from one conformation to another, as indicated in Figure 6.3 by the double arrows $\rightleftharpoons$. In physiological solutions, i.e. in the living organisms' cells, fluids and tissues, hexoses exist almost exclusively in the pyranose forms. For example, at 31°C,

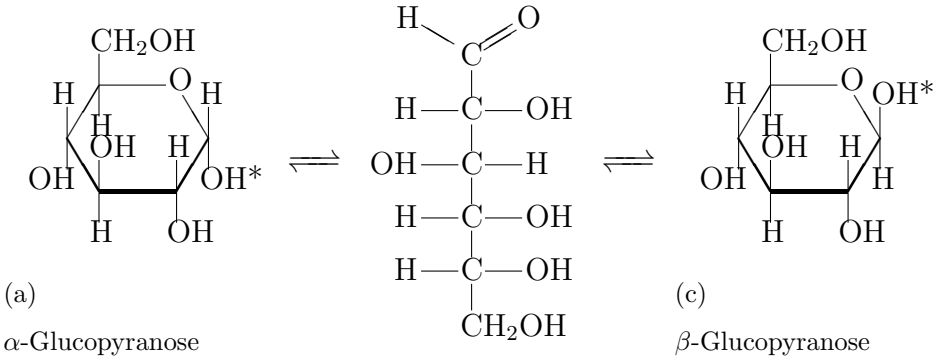


(a) Galactose (b) Glucose (c) Mannose

Figure 6.1: Structures of galactose, glucose and mannose. Carbon 5 hydroxyl group denoted by *.

(a) Carbonyl (b) Hydroxyl

Figure 6.2: Structures of carbonyl and hydroxyl groups



(a)
α-Glucopyranose

(b) Glucose

(c)
β-Glucopyranose

Figure 6.3: Glucose cyclization

Glucose exists in an equilibrium mixture of 64% β-Glucopyranose and 36% α-Glucopyranose, with only a tiny fraction in the open-chain form [57].

6.1.3 Hexose Binding

Binding proteins are characterized by a *binding site*: a cleft or groove in their structure where binding occurs. The molecule that binds to the protein is called a *ligand*. The binding process, known as *docking*, occurs in a key-lock fashion, where the binding site is tailored to accept and bind to this specific ligand.

Hexose binding proteins belong to diverse functional families that lack significant sequence similarity and, often, even structural similarity [65, 123]. Despite this fact, these proteins show high specificity to their hexose ligands. The 3-D conformation and the chemical properties of the few amino acids (also called residues) present at the binding site play a large role in determining the binding site's distinctive topology and biochemical properties; and hence the ligand type and the protein's functionality.

Proposed protein-sugar computational models are based, at least partially, on prior biochemical findings and knowledge [77, 117, 123, 127]. They incorporate different parts of these findings in predictive black-box models. No prior work has taken the opposite approach: given hexose binding sites data, what biochemical rules can we extract by just looking at the PDB features without incorporating other biochemical knowledge, and how do they compare to known rules? Hence we argue that there is a *need for a data-driven empirical validation of biochemical hexose-binding findings*.

Even though glucose is the second most abundant hexose, there is *no glucose-specific binding model*. It is thus interesting to *uncover the differences between glucose-specific and hexose general binding*.

In this chapter, we *build models for hexose and glucose binding, and let an expert compare the models to infer differential rules and features*. We use the Section 3.4 glucose and hexose dataset. The ED approach is not ILP-specific and can be applied to any classifier. We consider both ILP and SVMs. In doing so, we create the *first glucose-binding classifier*,

and perform an ILP-based *data-driven empirical validation of biochemical hexose-binding knowledge*.

6.2 Literature Review

Researchers have investigated protein-sugar binding sites for several years. Most of these attempts focused on galactose, since it is the most common hexose in biological processes.

6.2.1 Biochemical Approach

From the biochemical perspective, Rao *et al.* [105] fully characterized the architecture of galactose and mannose binding in Lectins, a major hexose-binding protein family. They identified four conserved residues that occupy identical positions independent of their sugar specificity, and interact with the hexose independent of its type. These invariant residues are Asp, Gly, Asn and an aromatic Phe/Tyr.

Later, Quiocho and Vyas [102] presented a review of the biochemical characteristics of carbohydrate binding sites and identified the planar polar residues (Asn, Asp, Gln, Glu, Arg) as the most frequently involved residues in hydrogen bonding. They also found that the aromatic residues Trp, Tyr, and Phe, as well as His, stack against the apolar surface of the sugar pyranose ring. Quiocho and Vyas also pinpointed the role of metal ions in determining substrate specificity and affinity. Ordered water molecules bound to protein surfaces are also involved in protein-ligand interaction [61].

Taroni *et al.* [127] analyzed the characteristic properties of sugar binding sites and described a residue propensity parameter that best discriminates sugar binding sites from other protein-surface patches. They also note that simple sugars typically have a hydrophilic side group which establishes hydrogen bonds and a hydrophobic core that is able to stack against aromatic residues. Sugar binding sites are thus neither strictly hydrophobic nor strictly hydrophilic, due to the dual nature of sugar docking. In fact, as García-Hernández *et al.* [50] showed, some polar groups in the protein-carbohydrate complex behave hydrophobically.

Furthermore, Zhang *et al.* [139] reported that the hydrogen bonds between the hexose ligand and certain amino acids in galactosyltransferases are crucial for the orientation of the ligand and the correct function of the protein.

6.2.2 Computational Approach

Some of this biochemical information has been used in computational work with the objective of accurately predicting sugar binding sites in proteins. Taroni *et al.* [127] devised a probability formula by combining individual attribute scores. Shionyu-Mitsuyama *et al.* [117] used atom type densities within binding sites to develop an algorithm for predicting carbohydrate binding. Chakrabarti *et al.* [23] modeled one glucose binding site and one galactose binding site by optimizing their binding affinity under geometric and folding free energy constraints. Other researchers formulated a signature for characterizing galactose binding sites based on geometric constraints, pyranose ring proximity and hydrogen bonding atoms [123, 124]. They implemented a 3D structure searching algorithm, COTRAN, to identify galactose binding sites. More recently, Malik and Ahmad [77] used a Neural Network to predict general carbohydrate as well as specific galactose binding sites.

On a broader scale, Gold and Jackson [52] compiled the SitesBase database of precalculated protein-ligand binding site similarities. They did this by performing an all-against-all geometric hashing over the Protein Data Bank (PDB). Given a binding site, SitesBase returns all entries with similar binding sites, ranked by a similarity score. Although the primary use of this database is to examine structural similarities between related binding sites, it can also provide evidence of functional similarity for unclassified binding sites.

Finally, segmentation and visualization techniques can be used to model protein cavities and binding-sites. Some approaches can detect protein surface-pockets of a given size [62, 136] and, when used as an input to a hexose-classifier, can present it with potential binding-sites for discrimination. Others techniques work on surface matching [28]: given a set of binding-sites in protein examples, they search for matching functional sites in other proteins.

6.3 Classifiers

Since the Expert Driven approach is not ILP-specific and can be applied to any classifier, we build both ILP and SVM models for each hexose and glucose classification problem.

6.3.1 SVM Model

The Support Vector Machines (SVM) classifier [129] requires a constant length feature vector representation. To generate this feature vector, we begin by subdividing the binding-site into 8 concentric layers. The first layer has a width of 3 Å and the subsequent 7 layers were 1 Å each. We then compute the cumulative number of atomic chemical properties (partial charge, hydrogen-bonding ability, hydrophobicity level) within each layer, as well as the cumulative number of residue groupings (Table 3.4). This scheme ensures a constant-length feature vector representation.

Since the classifier performance depends both on the sample size and the number of features, a rule of thumb is to provide at least ten times as many training samples per class as the number of features [59]. This is not the case with our small (yet exhaustive) dataset. Nevertheless, of all the different biochemical and geometrical features of a binding site, only some are essential for correct classification. Recognizing these features as part of a dimension-reduction step should improve the efficiency of our method [60].

Random Forests (RF) [15] is a classification algorithm based on multiple classification trees. RF provides measures of feature importance, and can be used as a feature selection tool [36]. Coupling RF feature selection with SVM classification tends to outperform SVM alone [26]. Its use is ideal in our case: RF feature selection is robust to noise, can be used when the number of features is much larger than the number of observations, incorporates feature interactions, and returns a direct feature importance measure.

RF feature selection improves the SVM classification on our data, as example Table 6.1 shows for Glucose. We report the number of *support vectors*, that is the number of data points supporting the SVM discriminating hyperplane. A smaller number of support vectors

reflects a better generalization [31]. By lowering the number of support vectors, RF-SVM increases generalization potential.

Table 6.1: Comparison of SVM's cross-validated performance on chemical and residue properties with and without RF feature selection over the glucose dataset

Property	RF	Feature Number	Error (%)	Sensitivity (%)	Specificity (%)	Support Vectors (%)
Charge	false	24	24.32	79.31	73.33	77.03
	true	5	14.86	86.21	84.44	44.59
Hydrogen	false	16	17.57	82.76	82.22	41.89
Bonding	true	3	14.86	82.76	86.67	47.30
Hydro-	false	24	16.22	72.41	91.11	65.57
phobicity	true	15	12.16	82.76	91.11	40.54
Residue	false	48	21.62	48.28	97.78	100.0
Grouping	true	19	09.46	93.10	88.89	41.89
Features	false	112	18.92	75.86	84.44	79.73
Combined	true	24	08.11	89.66	93.33	40.54

To infer relevant rules and features from RF-SVM, we first investigate the classification performance of each one of the biochemical features on its own (charge, hydrogen bonding, hydrophobicity, residue grouping). We then combine all features to form our final model (see Table 6.1). As an example of this technique, Figure 6.4 shows the charge features and their importance scores as returned by RF. We can see the importance of negatively charged atoms, especially in layer 3, the protein layer in contact with the docked hexose.

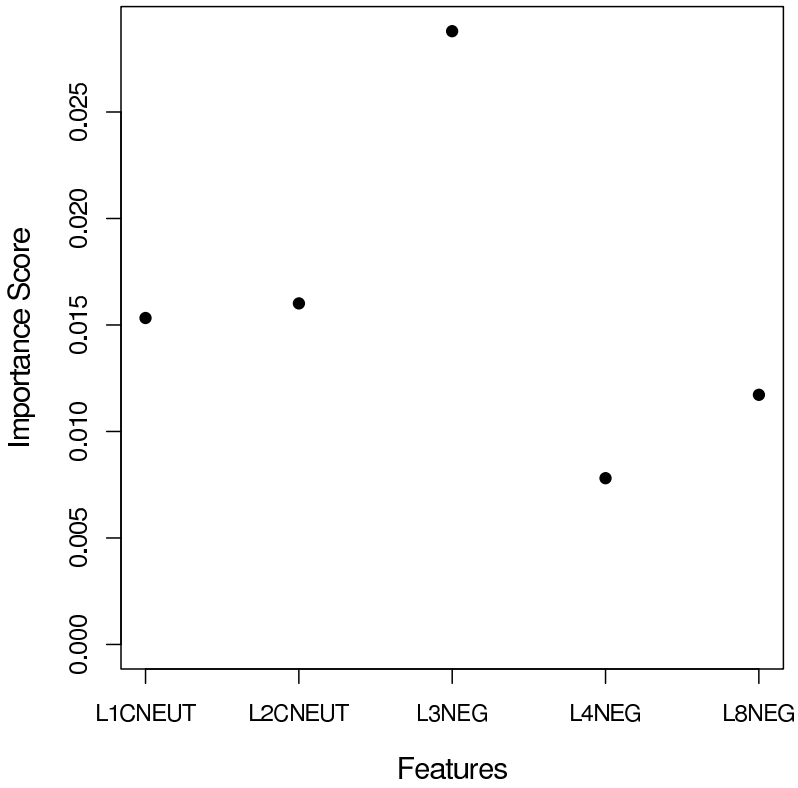


Figure 6.4: Importance of charge features according to RF over glucose dataset. NEUT stands for neutral, NEG for negative, and L# for the layer number.

6.3.2 ILP Model

After performing SVM runs, we note that the layers covering the first 5 Å, the subsequent 3 Å and the last 2 Å share several attributes. We thereby subdivide our binding-site sphere into 3 concentric layers, with layer width of 5 Å, 3 Å and 2 Å respectively. For each layer, we mine the total number of atoms in that layer and the cumulative number of each atomic property (charge, hydrogen-bonding, hydrophobicity).

We use the ILP engine Aleph [121] to learn first-order rules. The consequent of any rule is $bind(+site)$, where *site* is predicted to be a hexose binding site. No literal can contain terms pertaining to different binding sites. As a result, *site* is the same in all literals in a clause.

Literals describing individual PDB-atoms are of the form:

$$point(+site, -id, -X, -Y, -Z, -charge, -hbond, -hydro, -elem, -name) \tag{6.1}$$

where *site* is the binding site and *id* is the individual atom's unique identifier. X, Y, and Z specify the PDB-Cartesian coordinates of the atom. *charge* is the partial charge, *hbond* the hydrogen-bonding, and *hydro* the hydrophobicity. Lastly, *elem* and *name* refer to the atomic element and its name.

Clause bodies can also use distance literals:

$$dist(+site, +id, +id, \#distance, \#error)\,. \tag{6.2}$$

The *dist* predicate, depending on usage, either computes or checks the *distance* between two points. *site* is the binding site and the *id*s are two unique point identifiers. *distance* is their Euclidean distance apart and *error* the tolerated distance error, resulting in a matching interval of $distance \pm error$. We set *error* to 0.5 Å.

We want our rules to refer to properties of PDB-atoms, such as "an atom's name is ND1", or "an atom's charge is not positive". Syntactically we do this by relating PDB-atoms' variables to constants using "equal" and "not equal" literals:

$$equal(+setting, \#setting)\,, \tag{6.3}$$

$$not\_equal(+feature, \#feature)\,. \tag{6.4}$$

$feature$ is the atomic features *charge*, *hbond* and *hydro*. In addition to these atomic features, *setting* includes *elem* and *name*.

Aleph keeps learning rules until it has covered all the training positive set, and then it labels a test example as positive if *any* of the rules cover that example. This has been noted in previous publications to produce a tendency toward giving more false positives [33, 32]. To limit our false positives count, we restrict coverage to a maximum of 5 training-set negatives. Since our approach seeks to validate biological knowledge, we aim for high precision rules. Restricting negative rule coverage also biases generated rules towards high precision.

6.4 Expert Driven Differential Rules

After building the hexose and glucose models, we analyze the selected rules and features. We compare both models to uncover differential rules. Table 6.2 compares the equivalent rules found by the models.

Table 6.2: Rules and features of the glucose-specific and hexose-general models

Feature	Glucose	Hexose
Water and ions	x	x
Negative charge and carboxylate residue	x	x
Surface hydrogen bonding	x	x
Dual hydrophobic-hydrophilic	x	x
Aromatic residue docking		x

Glucoses being hexoses, most of the extracted rules and model features are similar. Their respective models confirm the relevance of water and ions in binding. Ordered water molecules and ions present at or near the binding cavity do play a role in determining substrate specificity and affinity [102, 61].

Both models also show a prominence of acidic residues (which have a negative partial charge) and of atoms with a negative partial charge. This is a known feature [102] and may be explained by the need to stabilize the dense hydrogen-bond network formed by the hexose hydroxyl groups. In fact, both models confirm the presence of hydrogen bonding atoms in direct contact with the docked hexose [139].

Hexose binding sites are neither hydrophobic nor hydrophilic, but rather exhibit a dual hydrophobic-hydrophilic nature where both antagonistic properties are involved in docking [127]. The hydrophilic region, composed of protruding hydroxyl groups, establishes hydrogen bonds. The hexose hydrophobic region is its pyranose ring, and it tends to stack over hydrophobic residues. Both models had rules pinpointing this uncommon dual nature.

Nevertheless, one key binding aspect was different. The pyranose ring, when stacking hydrophobically, tends to do so over an aromatic residue. These residues have large hydrophobic rings that the pyranose ring stacks over [116]. The hexose model highlighted this interaction.

The glucose model did not report this aspect. In fact, and unlike other hexoses, glucose stacks over an aromatic residue in most, but not all, binding sites [124]. This may be the reason why the glucose model finds hydrophobicity as the best discriminating atomic chemical property (see Table 6.1). Since the aromatic stacking is due to hydrophobic forces, and the hydrophobic interaction is present even in the absence of aromatic residues, the glucose model incorporated the ubiquitous hydrophobic feature as a whole instead of singling out its aromatic component.

This chapter has demonstrated that the Expert Driven approach to differential prediction, in conjunction with ILP or other machine learning algorithms, can uncover important knowledge about a domain. In the next chapter we show how a technical improvement to ILP can further improve the quality of learned knowledge in this same hexose-binding domain.

Chapter 7

Randomized and Domain-Dependent ProGolem

This chapter builds on the hexose work, where we *alter the recall selection of the ILP system ProGolem*. We establish that *randomized-recall ProGolem should be used as default* since it avoids data idiosyncrasies; and that recall selection, as well as other ProGolem settings, is domain-dependent. This chapter is based on a paper published in the journal *BMC Bioinformatics* [114].

7.1 Motivation

The hexose-binding ILP task is a highly non-determinate one. A residue can have multiple atoms, and the distance literal checks the distance from an atom to all other atoms. The complexity and size of the hypothesis space often presents computational challenges in search time which limit both the insight and predictive power of the rules found.

Top-down ILP systems, like Aleph [121], tend to use a "one-step lookahead" search strategy that assumes literals are conditionally independent given the target class. Even multi-step lookaheads and backtracking can not capture complex predicate dependencies. If the features are highly correlated, which is the case for hexose-binding, this results in the myopia effect [67], where a significant portion of the search resources is wasted searching very similar hypotheses, resulting in a poorer chance of finding good theories.

To address both these problems, we consider using ProGolem [82], a newly developed bottom-up ILP algorithm which is able to learn better than Aleph in highly non-determinate domains. It explores the search space lattice following a subsumption order relative to a

bottom clause, and is less prone to the myopia effect. For a review of ILP, Aleph and ProGolem, see Section 3.1.

We *theorize that a bottom-up approach is more suitable* in the case of non-determinate and correlated predicates. We note that ProGolem bounds the non-determinacy of non-ground predicates to low recall values, introducing placement bias by relying on the given order of ground terms. We argue that *randomizing the ProGolem selection of recall ground atoms would eliminate the placement bias.*

7.2 Non-Determinacy and Recall

In ILP, an example can have multiple instances from the same attribute. For illustration, a *person* has exactly one full $legalName$, two *parent* instances, and may have multiple *child* instances. Hence $legalName(Joe, X)$ has just one solution, $parent(Joe, Y)$ will have exactly two solutions, while $child(Joe, Z)$ can have any number of solutions. The number of possible solutions or answer substitutions of a given predicate is called its *non-determinacy.*

Determinate predicates may have at most one solution when their input arguments are instantiated. Hence $legalName$ is a determinate predicate, while *parent* and *children* are non-determinate.

In Prolog, background knowledge about predicates is encoded using mode declarations. Knowing and encoding the predicates non-determinacy helps ILP systems limit their search. The bound on the non-determinacy of a predicate is called its *recall.* The meaning of recall in ILP is not to be confused with the statistical measure of the same name (also called sensitivity) which is the fraction of correctly classified positive examples over all the positive examples.

For illustration, the mode declarations for our three predicates are:

$$\begin{aligned} &: -mode(1, legalName(+person, -name)). \\ &: -mode(2, parent(+person, -person)). \\ &: -mode(*, child(+person, -person)). \end{aligned} \tag{7.1}$$

Here, *recall* takes the values $1, 2$ and $*$ respectively. The value $*$ indicates the recall is not finite. We can also specify predicate recall values smaller than their non-determinacy. For instance, we can have:

$$:-mode(5, child(+person, -person)). \tag{7.2}$$

where we limit the number of children to five. If a person has more children, only the first five will be considered, the others will be ignored.

The concept of recall and non-determinacy is particularly relevant during saturation. Highly non-determinate predicates may result in an exponential growth in the number of bottom clause literals. Since ProGolem is a bottom-up algorithm starting the search from the bottom-clause, it is even more vulnerable to bottom-clause growth than Aleph. Non-determinacy makes ProGolem learning time exponential in the number of solutions considered, and it is often necessary to limit the recall to low values.

7.3 Altering ProGolem Recall

In highly non-determinate domains, such as hexose-binding, ProGolem bounds the recall to remain tractable. In other words, even if a predicate has multiple possible instantiations, only the first *recall* such instantiations are incorporated in ProGolem's bottom clause, and therefore in a hypothesis. A major drawback of this technique is that it introduces *placement bias*, by relying on the given order of ground terms. It is subject to data idiosyncrasies, discards many potentially useful ground atoms, and results in information loss.

As a remedy, we propose to *randomize ProGolem's selection of recall ground atoms.* This randomized recall approach considers all solutions first, out of which it randomly picks a number equal to *recall*; rather than the first *recall* atoms in the binding-site data representation.

We also notice that PDB-atoms closer to the binding center are more likely to influence binding [86]. We thus propose a domain-dependent recall approach where we order the

background PDB-atoms by their distance from the binding site center. The recall bound will only consider the *recall* atoms closest to the cavity centroid.

We thus consider three schemes. The first orders the PDB atoms according to their occurrence in the PDB file, which follows the protein primary sequence. The second scheme randomizes the order of the atoms in the background knowledge. The third scheme, *domain-dependent*, orders the atoms by their distance to the binding-site center. The three approaches respectively yield an accuracy of 59.4%, 68.8% and 74.4% (Table 7.1).

Table 7.1: 10-folds cross-validation predictive accuracies for ProGolem using different recall selection methods on the hexose dataset

	ProGolem recall selection method		
Fold	Primary sequence	Randomized	Domain-dependent
1	43.8%	56.3%	87.5%
2	62.5%	93.8%	78.5%
3	81.3%	87.5%	87.5%
4	56.3%	50.0%	43.8%
5	68.8%	68.8%	81.3%
6	37.5%	56.3%	81.3%
7	56.3%	62.5%	75.0%
8	68.8%	68.8%	81.3%
9	62.5%	81.3%	62.5%
10	56.3%	62.5%	68.8%
Mean	59.4%	68.8%	74.8%
Std Dev	12.6%	14.4%	13.4%

Sorting the binding-site atoms according to their distance from the binding center outperforms randomizing them, which in turn outperforms using their given PDB sequence order. Clever manipulations based on prior knowledge will have better results compared to default

settings. We thereby argue that *Randomized-ProGolem should be used as default* since it avoids data idiosyncrasies; and that recall selection, as well as other ProGolem settings, is domain-dependent.

7.4 Assessing Domain-Dependent ProGolem

Using the domain-dependent predicate ordering, we compare Aleph, ProGolem and RF-SVM. We consider both an *atom-only* representation (Table 3.5), and one augmented with *amino-acid* residue grouping information (Table 3.4). Table 7.2 shows the results.

7.4.1 ProGolem Performance

We notice that ProGolem performs better using the enhanced *amino acid* representation rather than *atom-only* (p-value = 0.029). However, the *amino acid* representation yields no statistically significant improvement in Aleph (p-value = 0.39). A possible explanation as to why ProGolem takes advantage of the amino acid representation more than Aleph is the myopia effect [67]. The myopia effect occurs because general-to-specific ILP systems, like Aleph, indirectly assume literals are conditionally independent given the target class. They refine the working hypothesis by adding one literal at a time, the one that maximizes a fitness function. If literals have a strong conditional dependency, any selected literal will roughly have the same score. Thus multiple literals need to be added before Aleph can determine which set is optimal. If the literals are highly non-determinate, as is our case, a significant portion of the search resources is wasted searching very similar hypotheses, which results in a poorer chance of finding good theories.

ProGolem outperforms Aleph for both representations. The differences in their predictive accuracies are statistically significant for both *atom-only* (p-value = 0.043) and *amino acid* (p-value = 0.004) representations, the latter being significant even at the 99% confidence level. This discrepancy is in part explained by ProGolem's global theory construction, which only constructs the final theory after all hypotheses have been generated rather than incrementally, on a per-example basis, as Aleph does.

Table 7.2: 10-folds cross-validation predictive accuracies for domain-dependent ProGolem, Aleph, and RF-SVM over the hexose dataset. The *1* besides Aleph and ProGolem stands for the atom-only representation and the *2* for the representation including amino acids. SVM uses a representation that includes amino acids.

	Learning algorithm				
Fold	Aleph 1	ProGolem 1	Aleph 2	ProGolem 2	RF-SVM
1	50.0%	75.0%	56.3%	75.0%	81.3%
2	68.8%	81.3%	68.8%	81.3%	87.5%
3	62.5%	68.8%	68.8%	93.8%	87.5%
4	50.0%	56.3%	68.8%	75.0%	75.0%
5	75.0%	81.3%	56.3%	81.3%	75.0%
6	68.8%	87.5%	81.3%	87.5%	87.5%
7	75.0%	81.3%	75.0%	81.3%	93.8%
8	93.8%	81.3%	75.0%	93.8%	87.5%
9	68.8%	75.0%	75.0%	81.3%	75.0%
10	56.3%	56.3%	87.5%	81.3%	62.5%
Mean	66.9%	74.4%	71.3%	83.2%	81.3%
Std Dev	13.2%	10.8%	9.8%	6.6%	9.3%

Finally, we compare ILP to RF-SVM. Despite *amino acid* ProGolem having a higher average accuracy and a lower standard deviation than SVM, the difference is not statistically significant (p-value = 0.52). More surprisingly, SVM does not significantly outperform *amino acid* Aleph (p-value = 0.057). SVM significantly outperforms both Aleph (p-value = 0.005) and ProGolem (p-value = 0.025) in the *atom-only* representation.

7.4.2 ProGolem Insight from Rules

ProGolem returned rules covering similar themes as Aleph (see Table 6.2). In addition, ProGolem mined two novel rules characterizing a hexose-binding site:

1. It contains a TYR residue whose CB and OH atoms are 5.6 ± 0.5 Å apart, a HIS residue whose ND1 atom is 8.9 ± 0.5 Å away from the binding center, and a TYR residue whose O atom is 9.8 ± 0.5 Å away from the binding center.
   [Positives covered = 6, Negatives covered = 0]

2. It contains CYS and LEU residues, and an ASP residue whose N and OD2 atoms are 4.6 ± 0.5 Å apart, and whose C atom is 7.6 ± 0.5 Å away from the binding center.
   [Positives covered = 18, Negatives covered = 0]

The first rule requires the presence of one or two TYR, and a HIS. This rule is thus describing a conformational representation of two or three aromatic residues around the binding-site center. It is interesting that this low-coverage rule may indeed be capturing the infrequent sandwich interaction, whereby two or more aromatic residues engage both faces of a hexose pyranose ring [13].

The second rule specifies CYS and LEU residues. Both have negative interface propensity measures and do not form hydrogen bonds with hexoses [127]. The interface propensity measure is defined as the logarithm of the ratio between a surface residue frequency at the sugar binding site, and the average frequency of any surface residue at the binding site. It is a measure that quantifies the disposition of amino acids to be in contact with the docked sugar. A residue with a negative propensity measure does not favor the sugar binding-site region since it is present there less frequently than average.

This rule covers 18 positive examples and no negative examples, and clearly specifies the presence of CYS and LEU as a discriminative factor for hexose-binding site recognition. This dependency over LEU and CYS is not previously identified in literature and merits further attention.

Chapter 8

BI-RADS Information Extraction

Our breast cancer dataset is mostly in a free-text format. Since ILP and most other machine learning classifiers operate on tabular data, an information extraction preprocessing step is required. This chapter presents *the first successful mammography information extraction application* from free-text mammogram records, as well as *the first breast tissue composition extractor*. We also confirm the application of this method on another dataset and in another language, namely creating *the first Portuguese mammography information extraction application*. This chapter is based on a paper presented at *ICDM-09 Workshops* [91], on another published in the *Journal of the American Medical Informatics Association (JAMIA)* [99], and on a third accepted at *BIBM-12* [88].

8.1 Problem Overview

The American College of Radiology (ACR) developed a specific lexicon to homogenize mammographic findings and reports: Breast Imaging Reporting and Data System (BI-RADS) [5]. The BI-RADS lexicon consists of 43 mammography descriptors organized in a hierarchy (Figure 8.1).

In radiology reports, these concepts are not uniformly described. Radiologists use different words to refer to the same concept. Some of these synonyms are identified in the lexicon (e.g. "equal density" and "isodense"), while others are not and need to be provided by experts (e.g. "oval" and "ovoid"). Some lexicon words are ambiguous, referring to more than one concept, or to no concept at all. The word "indistinct" may refer to the "indistinct

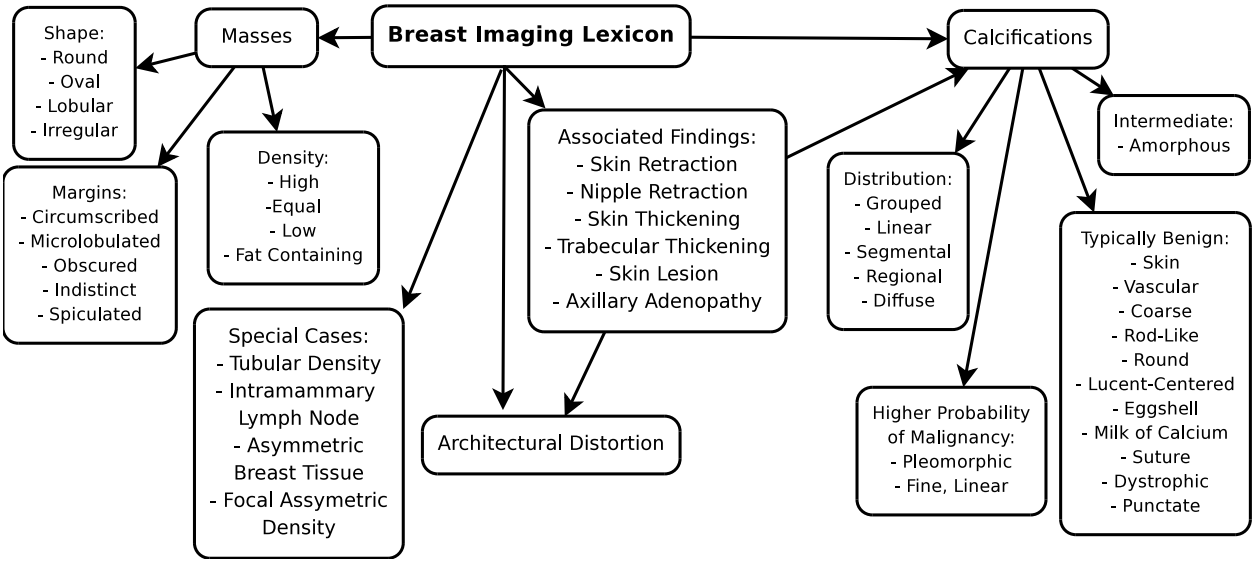


Figure 8.1: BI-RADS lexicon

margin" or to the "amorphous/indistinct calcification" concepts. Or it may be used in a non-mammography context, like "the image is blurred and indistinct".

Therefore, we cannot solely rely on the lexicon to map words and phrases in the text into concepts. A second level of complication arises from the presence of non-mammography medical concepts in the text. Negation presents a third substantial challenge, since pertinent negative observations often comprise the majority of the content in medical reports [24].

8.2 Literature Review

Only one prior study addresses BI-RADS information extraction from compliant radiology reports [17]. This research used a Linear Least Squares Fit to create a mapping between mammography report words-frequency and BI-RADS terms. It makes minimal use of lexical techniques and reports poor performance. However, several researchers tackled the similar problem of clinical information extraction from medical discharge summaries, discussed next.

8.2.1 Clinical Information Extraction

Most approaches to processing clinical reports heavily rely on natural language processing techniques. For instance, the MedLEE processor [45, 46] is capable of complex concept extraction in clinical reports. It first parses the text, using a semantic grammar to identify its structure. It then standardizes the semantic terms and maps them to a controlled vocabulary.

In parallel, the emergence of medical dictionaries emphasizes a phrase-match approach. The National Library of Medicine's Unified Medical Language System [71] (UMLS) compiles a large number of medical dictionaries and controlled vocabulary into a metathesaurus, which provides a comprehensive coverage of biomedical concepts. The UMLS metathesaurus was used to index concepts and perform information extraction on medical texts [7, 76]. Similar approaches have been used with more specialized terminology metathesauri, such as caTIES and SNOMED CT [22]. The BI-RADS lexicon can be seen as a metathesaurus for our task.

Finally, most clinical reports are dictated. They contain a high number of grammatically incorrect sentences, misspellings, errors in phraseology, transcription errors, acronyms and abbreviations. Very few of these abbreviations and acronyms can be found in a dictionary, and they are highly idiosyncratic to the domain and local practice [106]. For this reason, expert knowledge can contribute to effective data extraction.

8.2.2 Negation Detection in Clinical Documents

Negation presents another substantial challenge for information extraction from free text. In fact, pertinent negative observations often comprise the majority of the content in medical reports [24]. Fortunately, medical narrative is a sublanguage limited in its purpose, and its documents are lexically less ambiguous than unrestricted documents [109]. Clinical negations thus tend to be much more direct and straightforward, especially in radiology reports [84]. A very small set of negation words ("no", "not", "without", "denies") accounts for the large majority of clinical negations [25, 84].

Negation detection systems first identify propositions, or concepts, and then determine whether the concepts are negated. Basic negation detection methods are based on regular expression matching [25, 84]. More recent approaches add grammatical parsing [58], triggers [51] and recursion [107].

8.3 Mammography Feature Extraction Algorithm

In order to map words and phrases in the text into mammography concepts, we supplement the BI-RADS lexicon by a semantic grammar that maps the underlying BI-RADS categories into well-defined semantic patterns. Our approach has three main modules (Fig. 8.2). Given the free-text BI-RADS reports, it first applies a syntax preprocessor. Then the semantic parser maps subsentences to concepts. Finally a lexical scanner detects negated concepts and outputs the tabulated BI-RADS features.

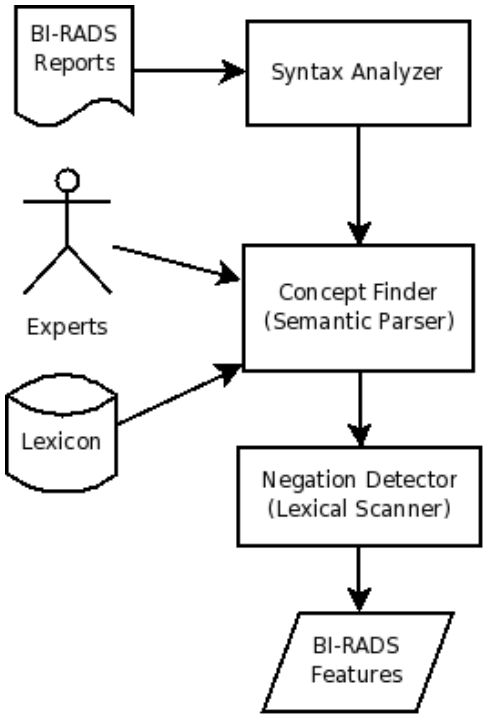


Figure 8.2: BI-RADS extraction algorithm flowchart

8.3.1 Syntax Analyzer

The first module in our system is a preprocessing step that performs syntactic analysis. Since BI-RADS concepts do not cross sentence boundaries, we process the reports by individual sentences. We then remove all remaining punctuation. We keep stop words because some of them are used in the negation detection phase.

8.3.2 Concept Finder

The concept finder module takes the syntactic token (a sentence) and applies grammar rules to search for concepts. Due to different word forms and misspellings, we use stem-words and ease our matching constraints. We formulate the rules as a context free grammar, and express them using Perl's pattern matching capacities [133].

For each BI-RADS concept, we first start with a rule that is solely based on the lexicon. This rule is then iteratively refined by an expert radiologist, who monitors the rule's performance over the training set. The expert establishes the order and scope of a rule, and provides domain synonyms, acronyms and idiosyncrasies.

As an example, the lexicon specifies the word "regional" to represent the "regional" distribution concept. The initial rule, searching for sentences with the word "regional", returns many false positives. Experts refine the rule to "regional not followed by medical or hospital".

8.3.3 Negation Detector

Once the semantic grammar detects a concept occurrence, it hands the subsentence token to the negation detection module. The negation detection module is a lexical scanner that searches for negation signals using regular expressions. It analyzes their negation scope to determine if they apply over the concept.

Following the approach of [51], we identify adverbial ("not", if not preceded by "where") and intra-phrase ("no", "without") negation triggers. Similar to previous findings [84], we find that negation triggers usually precede the concepts they act upon. In addition, since

our approach maps a concept to a subsentence, the negation trigger may appear within the concept's underlying indexed text structure. For instance, the word "mass" followed by "oval" within 5 words, is a rule for the "oval shape" concept. The subsentence "mass is not oval" is a negation within the concept.

We also note that there may be several words between the negation trigger and the concept it negates, and a single trigger may negate several concepts. The maximum degree of word separation between a trigger and its concept, referred to as the negation scope, differs among concepts. Accurate analysis of scope may involve lexical, syntactic, or even semantic analysis. We establish each concept's negation scope by counting and looking at a subset of the trigger's hits over the unlabeled training set. Starting with a high scope, we assess the number of false positives we get. With smaller scopes, we can assess the number of false negatives. We choose the scope that minimizes the error ratio.

Since we treat each concept occurrence individually, we can correctly detect a concept in a sentence containing both the concept and its negation. We hence avoid the pitfall of erroneously rejecting a concept encountered by [25], who negated the entire concept if a single instance of that concept was negated.

While analyzing negation errors, [84] reported errors caused by double negatives. We address this issue using the same approach to detect negation triggers. We identify a set of double-negation triggers which, when coupled with negation triggers, deactivate them. These signals are: "change", "all", "correlation", "differ" and "other". Therefore "there is no change in rounded density" does not negate the concept "round shape".

8.3.4 Handling Latent Concepts

Multiple latent concepts may exist in a given report. For instance, our mammography reports often contain ultrasound concepts. Ultrasound and mammography concepts can have common underlying words, thus the need to discriminate them. A "round mass" is a BI-RADS feature, while a "round hypoechoic mass" is an ultrasound feature. We use an ultrasound lexicon, composed of the concepts "echoic" and "sonogram" and apply the same

approach (Fig. 8.2) to detect ultrasound concepts. We require that a BI-RADS concept not share common subsentences with an ultrasound concept. Our method is thus able to handle multiple latent concepts within the text.

8.4 BI-RADS Features Extractor

We train our BI-RADS features extractor on the original 146,972 mammograms UCSF dataset (see Section 3.2). To test our method, we compare our algorithm's results to manual information extraction performed by radiologists. Our testing set consists of 100 records from the database that a radiologist on our team manually indexed in 1999 [17].

8.4.1 BI-RADS Extractor Methodology

Each record has a Boolean feature vector of 43 elements representing the BI-RADS lexicon categories (see Fig. 8.1). The information extraction task is to correctly populate the $43 \times 100 = 4300$ elements matrix by assigning an element to 1 if its corresponding BI-RADS feature is present in the report, and to 0 otherwise. The manual method extracted a total of 203 BI-RADS features, leaving 4097 empty slots.

The algorithm, on the other hand, extracts a total of 216 BI-RADS features, out of which 188 are in agreement with the manual extraction. In 43 cases, only one of the methods claims the presence of a BI-RADS feature. Upon review of these disparate results, a radiologist determined that our algorithm correctly classified 28 cases while the manual method correctly classified 15.

Clearly the manual method, applied in 1999, does not constitute ground truth. In fact, correctly labeling a text corpus is complicated enough that even experts need several passes to reduce labeling errors [41]. Due to the high labeling cost, in practice one must rely on the imperfect judgments of experts [118]. Since time spent cleaning labels is often not as effective as time spent labeling extra samples [69], our reviewing radiologist reexamined only the diverging cases.

We consider as ground truth the features that both computational and manual methods agree on, in addition to the relabeling of diverging cases by experts. This approach is likely underestimating the number of true features. The omission error of a method is bounded by the number of diverging cases correctly labeled by the other method. We assume that the classifier and the labelers make errors independently, since humans and computers generally classify samples using different methodologies. We use Lam and Stork's method of handling noisy labels [69]: we treat the classification differences between the two methods as apparent errors, and the classification differences between each method and ground truth as labeling errors. We factor both error terms to get the true classification errors and the confusion matrices for both our algorithm and the manual method (Table 8.1).

Table 8.1: Automated and manual extraction, 1^{st} run

		Actual	
Method	Predicted	Feature present	Feature absent
Automated	Feature present	211	5
	Feature absent	10	4074
Manual	Feature present	198	5
	Feature absent	23	4074

To compute test statistics, we treat the present features as positives and the absent features as negatives. Our data being highly skewed, we employ precision-recall analysis instead of accuracy. For the double-blind run, the manual method achieves a 97.5% precision, a 89.6% recall rates and a 0.93 F_1-score. Our algorithm achieves a much better recall (95.5%) and F_1-score (0.97) for a similar precision (97.7%). It correctly classifies 65.1% of the disputed cases.

To compare both methods, we use the probabilistic interpretation of precision, recall and F-score [53]. Using a Laplace prior, the probability that the computational method is

superior to the manual method is 97.6%. Our result is statistically significant at the 5% level (p-value = 0.024).

As in most clinical data, false negative mammograms are critical and often more costly than false positive ones [100]. Many technical or human errors cause missed or delayed diagnosis of breast cancer. Among the several reasons are observer error, unreasonable diagnostic evaluation, and problems in communication [16]. Therefore, it is notable that the main gain of our algorithm is in recall, by achieving low false negative counts. The algorithm's recall rate of 95.5% is higher than the manual method's 89.6% and the Linear Least Squares Fit method's reported 35.4% recall rate [17].

8.4.2 BI-RADS Extractor Final Model

Before the first run, we only adjusted the algorithm using unlabeled data. After performing the first run on labeled data, the experts suggested slight changes to some of the rules. We consider this modified version our final algorithm and use it for extracting terms from the UCSF database. This approach can be viewed as utilizing both labeled and unlabeled data to modify the algorithm [94]. Using the final version of the algorithm, we perform a second run over the test data (Table 8.2). Note that the test set is no longer a valid test set, since we looked at it to modify the algorithm. We are showing the results as a confirmation step, due to the lack of ground truth and the small number of labeled data.

During the second run, the algorithm correctly classifies some of its previous mismatches, dropping its false positive and false negative counts. It now achieves a precision of 99.1%, a recall of 98.2% and an F_1-score of 0.99. In addition, the algorithm discovers two more previously unrecognized true positives, which increases the manual method's false negative count.

8.4.3 Cross-Institution Portability

After training our parser on the UCSF data, we use it to extract BI-RADS features from mammography records at the Marshfield Clinic. We first validate our algorithm on 71

Table 8.2: Automated and manual extraction, 2^{nd} run

		Actual	
Method	Predicted	Feature present	Feature absent
Automated	Feature present	219	2
	Feature absent	4	4075
Manual	Feature present	198	5
	Feature absent	25	4072

reports manually annotated by a trained non-radiologist collaborator. A radiologist reviewed and re-annotated the diverging cases. Our algorithm achieves 97.9% precision and 95.9% recall, significantly outperforming the manual method. This result suggests cross-institution portability of our software.

8.5 Portuguese BI-RADS Features Extractor

To test the applicability of our approach to other languages, we apply our iterative method to Portuguese, resulting in the first Portuguese BI-RADS feature extractor. Our annotated dataset comes from the Centro Hospitalar São João in Porto, Portugal. It consists of 153 patients each of whom has one basic screening and one detailed diagnostic text report.

8.5.1 Portuguese Extractor Methodology

In order to build our parser, our first step was to translate the BI-RADS lexicon to Portuguese. This was done with the help of a specialist. We then proceeded in a similar manner as described previously. We built a dictionary of synonyms for every BI-RADS term. Using an iterative process, we supplemented this list using expert knowledge to differentiate between different uses of the same word, to gauge the proximity of the words of a multi-word concept, and to capture medical wording practices and idiosyncrasies. We perform stemming

and group words in the same concept if they are synonyms or typos. After detecting a concept, we proceed to the treatment of negations. Following [51], we identify a set of negation triggers: "não" (not) when not preceded by "onde" (where), "sem" (without), and "nem" (nor).

The evaluation of the parser was done in 3 phases. In the first phase, we only used the translated terms to extract the features. After reviewing the results with the specialist, we augmented our parser with synonyms and fine-tuned the word proximity for multi-words concepts. We performed this process of iterative expert knowledge incorporation over two iterations, constituting phases 2 and 3 of our analysis. The algorithmic performance also prompted the radiologist to update her own classification, since the parser was discovering BI-RADS features that she overlooked in her manual annotation.

Tables 8.3 and 8.4 show the total number of features extracted by the parser and the radiologist during the 3 different phases, for both the screening and the diagnostic mammograms. We group the extracted features according to the BI-RADS hierarchy (Figure 8.1).

8.5.2 Portuguese Extractor Results

During the first phase of evaluation, and using the screening mammogram reports, the parser extracted 44 features while the radiologist extracted 66. Out of 92 distinct extracted features, both methods had 18 features in common (20%), and disagreed on the remaining 74. Using the diagnostic mammography reports, the parser returned 71 features, and the manual method 122. Out of 160 distinct extracted features, both methods agreed on 33 (21%), and disagreed on the remaining 127. This was a double-blind experiment, where the parser and radiologist were not influenced by each other.

We discussed the first set of results with the radiologist, reviewing the parser's vocabulary. We refined its internal rules accordingly, and parsed the texts again. On the screening reports, the parser thus returned 87 features. Out of 99 distinct extracted features, the parser and radiologist had 54 cases in common (54%), a substantial improvement from the

Table 8.3: Number of attributes extracted from the screening mammograms, grouped by category

Concept		1st phase	2nd phase		3rd phase
	Radiologist	Parser	Parser	Radiologist	Parser
Shape	8	11	16	8	14
Margin	15	12	26	15	20
Density	0	4	2	0	1
Calc. Morphology	5	4	6	4	6
Calc. Distribution	8	2	8	8	9
Special Cases	9	7	8	7	7
Associated Findings	21	4	21	21	22
Total	66	44	87	63	80

first phase. For the diagnostic reports, the parser extracts 129 features. From a total of 146 distinct extracted features, 107 are agreements (73%) while 37 are disagreements, a significant improvement related to the first phase.

After phase 2, we performed a second round of parser fine-tuning. The radiologist too revised her annotations, removing 3 features from the screening matrix and adding 5 to the diagnostic matrix. Clearly the first manual extraction did not constitute ground truth. In fact, correctly labeling a text corpus is complicated enough that even experts need several passes to reduce labeling errors [41]. We can not assert what is ground truth, nor the actual number of features truly present in the text. Hence, we assume that the cases that both computational and manual methods agree upon are correctly classified, and we focus our attention on analyzing and re-labeling the disputed cases.

In the last phase, considering the screening reports, the parser returns 80 features and the radiologist 63. The two methods agreed on 59 extracted features, and differed on 25. Re-labeling the latter cases, the parser correctly classifies 14, versus 11 for the manual method.

Table 8.4: Number of attributes extracted from the diagnostic mammograms, grouped by category

Concept		1st phase	2nd phase		3rd phase
	Radiologist	Parser	Parser	Radiologist	Parser
Shape	3	1	4	3	3
Margin	22	18	24	22	24
Density	21	4	21	21	22
Calc. Morphology	9	9	13	10	14
Calc. Distribution	13	30	14	13	12
Special Cases	11	1	18	15	15
Associated Findings	43	8	35	43	36
Total	122	71	129	127	126

For the diagnostic reports, the parser and the radiologist respectively extract 126 and 127 features, forming 115 agreements and 23 disagreements. Our program correctly classified 11 of the disputed cases, while the radiologist got 12.

Combining both data subsets together, we can see that our method extracted 206 features, 174 of which are in accordance with manual extraction (84.5%). It extracted 32 features that the expert did not, while the radiologist had 16 extra features. Out of these 48 disputed cases, the parser edges the radiologist by correctly classifying 25 (52.1%). The parser is thus able to discover features missed or misclassified by the radiologist, and exhibits a similar performance. In fact, the parser returns 96.6% precision and 92.6% recall.

Figure 8.3 summarizes the improvements of the parser during the three phases of the experiment, in terms of concordant and discordant extracted features. Each phase is represented by four bars. The first two bars correspond to the screening reports while the next two correspond to the diagnostic reports. Taken in pairs, the left bar (Screening-C and Diagnostic-C) reports the number of concordances between the parser and the radiologist,

while the right bar (Screening-D and Diagnostic-D) reports the discordances, features that were either extracted by the parser or by the radiologist but not by both. For the diagnostic reports, we observe a drastic improvement between the first and second phases, and an additional slight improvement by the third phase. For the screening reports, the improvement is not so pronounced, because this type of reports is less thorough and detects less BI-RADS features.

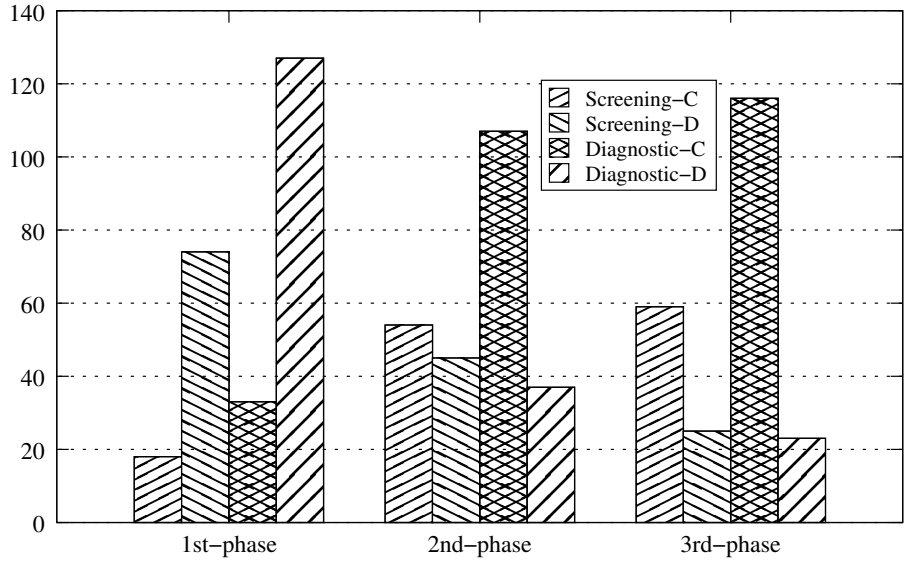


Figure 8.3: Number of concordant and discordant extracted features by the parser and the manual methods, over the three phases and both data subsets

8.6 Breast Tissue Composition Extractor

Breast tissue composition is an important component of the radiological evaluation of the breast for two reasons. First, dense fibroglandular tissue is a risk factor for breast cancer [14]. Second, this dense tissue decreases mammographic sensitivity in detecting breast cancer [21]. For these reasons, mammography reports typically contain a description of the overall tissue composition of the breast.

Although we did not originally consider breast tissue composition as part of the BI-RADS lexicon (Figure 8.1), the latter divides breast tissue density into four categories: 1

(predominantly fat), 2 (scattered fibroglandular densities), 3 (heterogeneously dense), and 4 (extremely dense) [5]. These standard categories help to minimize ambiguity in mammography reporting and also facilitate large-scale clinical studies of breast cancer, which must control for known risk factors like breast density. Reliable, standardized information on breast tissue composition could play an important role in the development of classification systems for the early detection of malignancy.

Unfortunately, breast composition information is typically not reported in coded form, and there is no automated method for extracting it from free text. We therefore apply our mammography information extraction approach to breast composition, resulting in the first automated method for detecting and extracting the breast density assessments from free-text mammography reports.

8.6.1 Breast Tissue Composition Extractor Methodology

For training our parser, we use the UCSF non-annotated training dataset described previously, in addition to the 34,489 reports Stanford RADTF (RADiology Teaching File) database [37]. We test the resulting classifier on two independent test sets, 500 annotated reports from the Stanford corpus (which were held out during the rule-construction phase), and 100 annotated reports from the Marshfield Clinic.

We apply our BI-RADS features extraction approach to retrieve breast densities, and augment it using the set of patterns observed on the Stanford data. Incorporating this expert knowledge into the iterative concept finder, we generate multiple pattern matching and regular expression rules, that automatically detect and extract BI-RADS breast density classes. Figure 8.4 shows the resulting classification criteria for each BI-RADS breast tissue composition class.

8.6.2 Breast Tissue Composition Extractor Results

We test our algorithm on the annotated Stanford and Marshfield testing sets. Two different radiologists reviewed the reports to establish a gold standard for comparison. We

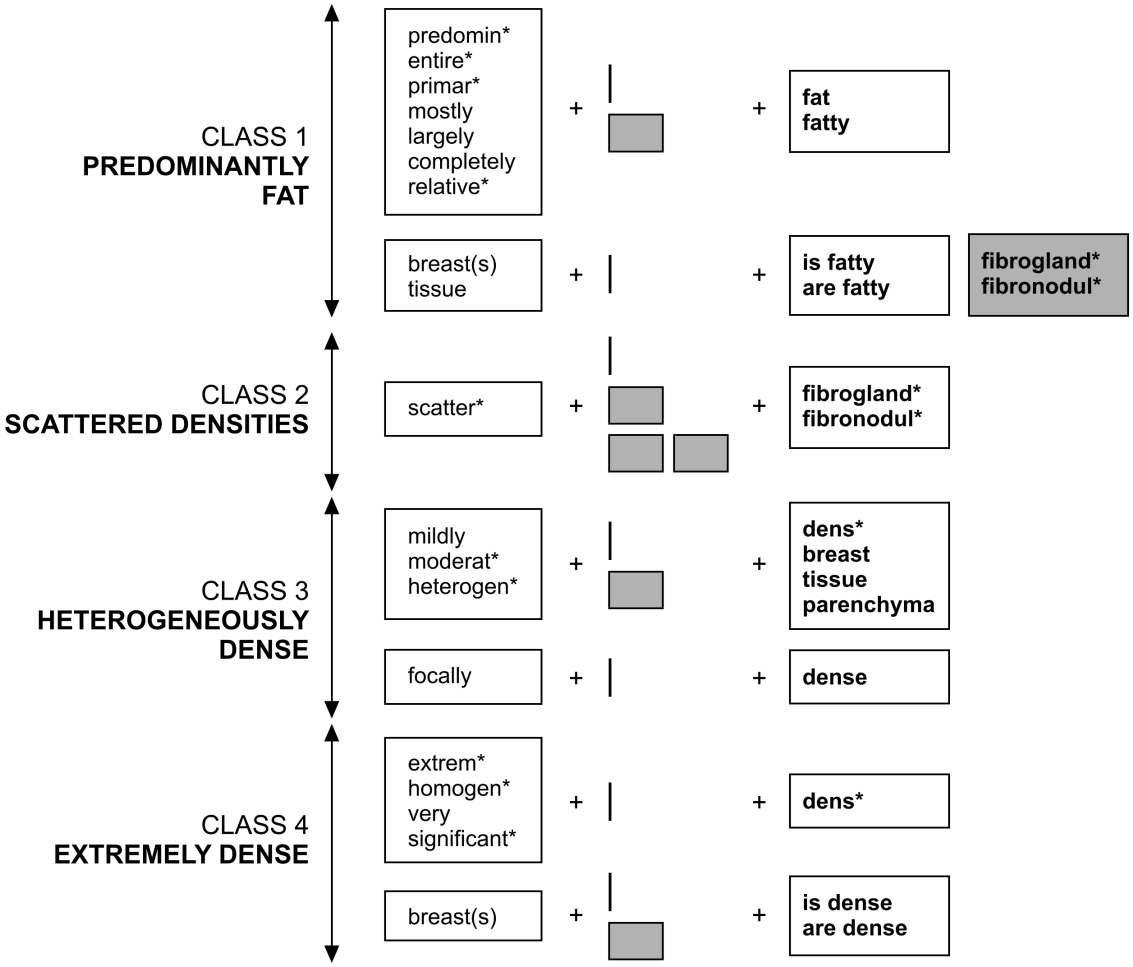


Figure 8.4: Rules used to assign reports to different BI-RADS tissue composition classes. White rectangles represent sets of words that must be present at a given location to fulfill the rule. Gray rectangles represent words that cannot be present at a location for the rule to be fulfilled. Small gray boxes represent unspecified words. The asterisk (*) is used to denote multiple possible word endings.

classify every mammography record as having a breast density category $1-4$, or "no descriptors"(Table 8.5).

Our algorithm correctly classified 499/500 (99.8%) reports from the Stanford dataset and 99/100 (99%) reports from the Marshfield Clinic dataset. On the Stanford data, the

Table 8.5: System performance results on the Stanford and Marshfield testing sets

Dataset	Records with descriptors present	Records with no descriptors	Correctly classified records	Total
Stanford	497	3	499	500
Marshfield	73	27	99	100

only wrongly classified report contained the description "bilateral breasts re-demonstrate dense glandular tissue", which the radiologist classified as class 4 and the algorithm as "no descriptors". On the Marshfield side, the radiologist assigned category 2 to "the right breast shows fibroglandular tissue which is finely nodular and strandlike", while the algorithm considered it as "no descriptors". Including "fibroglandular tissue" in the rules for class 2 led to many false positives for that class, and therefore we did not change the rules to accommodate this special case.

In conclusion, we have created an algorithm that automatically processes unstructured, free-text mammography reports and reliably extracts BI-RADS features and breast composition. This method could facilitate research and policy analysis by enabling investigators to efficiently mine large collections of mammography reports. Our approach can be applied to extract different mammography features, has a robust cross-institution portability, and can extend to other languages.

Chapter 9

Conclusion

Building differential prediction classifiers is a new and open research field. Standard classifiers may exhibit significant differences in performance over parts of the input space. Modeling this differential prediction behavior and building classifiers that maximize differential prediction over specific data subsets is an interesting research problem with several real-world applications.

9.1 Summary

This work constitutes the first attempt at learning differential predictive rules, and at extending differential prediction to relational datasets.

We start with a motivation for the task of differential prediction, followed by a review of prior differential work. Differential prediction originated in psychology to assess fairness of cognitive and educational tests. Considered an indicator of test bias, it is detected using logistic regression. The classification literature has extended the differential prediction concept to differences in predicted performance when an instance is classified into one condition rather than into another. Known as *uplift modeling* in marketing, it is modeled using various classifiers. We also review the use of rules for differential prediction, and propose a novel formulation of differential predictive rules.

Before introducing our attempts to address the multi-relational differential prediction problem, we cover the necessary background. We present an overview of Inductive Logic Programming (ILP) and the two ILP systems we use, Aleph and ProGolem. Our main

application is to uncover age-specific breast cancer stage differential prediction rules. Our secondary application is to infer differences between specific glucose and general hexose binding. We also consider a synthetic Michalski-trains dataset. We explain the collection and preprocessing steps pertaining to the datasets. We also review the methodologies to compare differential prediction results. We use the area under the precision-recall curves over the predicted rules if ground-truth rules are known. Otherwise, we use uplift curves over the classified instances.

We explore several methods to learn differential rules in a two-class two-strata system. The *Model Filtering* (MF) approach builds a rule-based model on the target stratum, and then selects rules that exhibit a differential performance on the other stratum. The *Differential Prediction Search* (DPS) method alters the search space to consider both strata while scoring rules according to their differential prediction score. Both methods are automated. The basic *Expert Driven* (ED) approach constructs a model on each dataset, and lets an expert compare them and infer differential rules. ED is non-automated and can be used with non-rule-learners.

We apply and compare the MF and DPS methods over the synthetic Michalski-trains dataset, and over the mammography dataset. Our results show that, for large and noisy data, which is what most real world applications are, DPS is more appropriate. For small and non-noisy data, MF outperforms DPS. Our methods, especially DPS, inferred rules and models that experts judged plausible and interesting. We also augment a Bayes Net with differential rules for risk prediction, forming a Logical Differential Prediction Bayes Net (LDP-BN), and observe a significant performance increase. I thus confirm my thesis statement, establishing that ILP-based differential relational classifiers can effectively propose rules that apply to a given multi-relational data subset, maximize performance differences over a stratified dataset, and offer significant insight into the underlying domain.

For illustration, we use the ED approach to infer differences between specific glucose and general hexose binding. We apply this method to ILP and SVMs classifiers. In doing so,

we devise the first glucose-binding classifier, empirically validate biochemical hexose-binding knowledge, and infer new hexose-binding and breast-cancer dependencies.

So far we have been using Aleph, a top-down ILP system. ProGolem, a bottom-up ILP algorithm is more suitable in the case of non-determinate and correlated predicates, which is the case for the hexose dataset. We alter the ProGolem recall selection, and further improve the quality of learned knowledge in the hexose-binding domain. We consider three recall selection schemes: default ordering, randomized ordering, and domain-dependent ordering. We establish that randomized-recall ProGolem should be used as default since it avoids data idiosyncrasies; and that recall selection, as well as other ProGolem settings, is domain-dependent.

Our breast cancer dataset is mostly in a free-text format. Since ILP and most other machine learning classifiers operate on tabular data, an information extraction preprocessing step is required. Our final contribution is to present an information extraction method for free-text mammogram reports. It resulted in the first successful mammography information extraction application, as well as the first breast tissue composition extractor. We also confirm the application of this method on another dataset and in another language, namely creating the first Portuguese mammography information extraction application.

9.2 Future Work

This work can be extended in several directions. We focused on addressing the two-class two-strata differential rule prediction problem. A natural extension is to consider multi-class and multi-strata problems. One may try reducing the K-strata problem to K 2-strata subproblems. Repeating K times, we keep one stratum and collapse the others together, creating a 2-strata one-versus-all subproblem. For each subproblem, we extract differential predicting rules pertaining to the specified stratum.

A different approach would use a differential-prediction-sensitive scoring function that applies to multiple strata. Finding a suitable function requires more thought and research. A possible exploration direction is f-divergence functions.

Second, we note that LDP-BN rules are learned for their differential predictive potential, separately from the Bayes Net. Integrating the differential rules identification and the Bayesian Network construction into a global optimization framework may result in a better performance [32].

One can argue that uplift modeling is a special case of differential prediction, where the score to maximize is the uplift score. We can implement the uplift function within ILP, creating a logical relational uplift model.

For our mammography information extraction system, we can use our rules to extend the Knowtator general-purpose text annotation tool [95] to include mammography. Our parser can also be refined by adding a syntactic parser and following the approach used by [132].

LIST OF REFERENCES

[1] S. Aebi and M. Castiglione. The enigma of young age. *Ann. Oncol.*, 17(10):1475–1477, 2006.

[2] C. J. Allegra, D. R. Aberle, P. Ganschow, S. M. Hahn, C. N. Lee, S. Millon-Underwood, M. C. Pike, S. Reed, A. F. Saftlas, S. A. Scarvalone, A. M. Schwartz, C. Slomski, G. Yothers, and R. Zon. National Institutes of Health State-of-the-Science Conference Statement: Diagnosis and Management of Ductal Carcinoma In Situ, September 22–24, 2009. *J. Natl. Cancer Inst.*, 102(3):161–169, 2010.

[3] American Cancer Society. *Breast Cancer Facts & Figures 2009-2010.* American Cancer Society, Atlanta, USA, 2009.

[4] American Cancer Society. *Cancer Facts & Figures 2009.* American Cancer Society, Atlanta, USA, 2009.

[5] American College of Radiology, Reston, VA, USA. *Breast Imaging Reporting and Data System (BI-RADS^{TM})*, 3rd edition, 1998.

[6] American Educational Research Association/American Psychological Association/National Council on Measurement in Education. *The Standards for Educational and Psychological Testing*, 1999.

[7] A. R. Aronson. Effective mapping of biomedical text to the UMLS metathesaurus: The MetaMap program. In *Proc. of the American Medical Informatics Association Symposium*, pages 17–21, Washington, DC, 2001.

[8] M. Ayvaci, O. Alagoz, J. Chhatwal, A. del Rio, E. Sickles, H. Nassif, K. Kerlikowske, and E. Burnside. Predicting invasive breast cancer versus DCIS in different age groups. Submitted to PLoS ONE.

[9] S. C. Bagley and R. B. Altman. Characterizing the microenvironment surrounding protein sites. *Protein Science*, 4(4):622–635, 1995.

[10] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The protein data bank. *Nucleic Acids Research*, 28(1):235–242, 2000.

[11] M. J. Betts and R. B. Russell. Amino acid properties and consequences of substitutions. In M. R. Barnes and I. C. Gray, editors, *Bioinformatics for Geneticists*, pages 289–316. John Wiley & Sons, West Sussex, UK, 2003.

[12] L. Bobadilla, F. Nino, and G. Narasimhan. Predicting and characterizing metal-binding sites using Support Vector Machines. In *Proceedings of the International Conference on Bioinformatics and Applications*, pages 307–318, Fort Lauderdale, FL, 2004.

[13] A. B. Boraston, D. N. Bolam, H. J. Gilbert, and G. J. Davies. Carbohydrate-binding modules: fine-tuning polysaccharide recognition. *Biochem. J.*, 382:769–781, 2004.

[14] N. F. Boyd, L. J. Martin, M. Bronskill, M. J. Yaffe, N. Duric, and S. Minkin. Breast tissue composition and susceptibility to breast cancer. *Journal of the National Cancer Institute*, 102(16):1224–1237, 2010.

[15] L. Breiman. Random Forests. *Machine Learning*, 45(1):5–32, 2001.

[16] R. J. Brenner. False-negative mammograms: Medical, legal, and risk management implications. *Radiol. Clin. N. Am.*, 38(4):741–757, 2000.

[17] B. Burnside, H. Strasberg, and D. Rubin. Automated indexing of mammography reports using linear least squares fit. In *Proc. of the 14th International Congress and Exhibition on Computer Assisted Radiology and Surgery*, pages 449–454, San Francisco, CA, 2000.

[18] E. S. Burnside, J. Davis, J. Chhatwal, O. Alagoz, M. J. Lindstrom, B. M. Geller, B. Littenberg, K. A. Shaffer, C. E. Kahn, and D. Page. Probabilistic computer model developed from clinical data in national mammography database format to classify mammographic findings. *Radiology*, 251:663–672, 2009.

[19] E. S. Burnside, D. L. Rubin, and R. D. Shachter. Using a Bayesian network to predict the probability and type of breast cancer represented by microcalcifications on mammography. *Stud Health Technol Inform*, 107(Pt 1):13–17, 2004.

[20] F. A. Carey. *Organic Chemistry*. McGraw-Hill, 5th edition, 2003.

[21] P. A. Carney, D. L. Miglioretti, B. C. Yankaskas, K. Kerlikowske, R. Rosenberg, C. M. Rutter, B. M. Geller, L. A. Abraham, S. H. Taplin, M. Dignan, G. Cutter, and R. Ballard-Barbash. Individual and combined effects of age, breast density, and hormone replacement therapy use on the accuracy of screening mammography. *Annals of Internal Medicine*, 138(3):168–175, 2003.

[22] D. Carrell, D. Miglioretti, and R. Smith-Bindman. Coding free text radiology reports using the cancer text information extraction system (caTIES). In *American Medical Informatics Association Annual Symposium Proceedings*, page 889, Chicago, IL, 2007.

[23] R. Chakrabarti, A. M. Klibanov, and R. A. Friesner. Computational prediction of native protein ligand-binding and enzyme active site sequences. *Proceedings of the National Academy of Sciences of the United States of America*, 102(29):10153–10158, 2005.

[24] W. W. Chapman, W. Bridewell, P. Hanbury, G. F. Cooper, and B. G. Buchanan. Evaluation of negation phrases in narrative clinical reports. In *Proc. of the American Medical Informatics Association Symposium*, pages 105–109, Washington, DC, 2001.

[25] W. W. Chapman, W. Bridewell, P. Hanbury, G. F. Cooper, and B. G. Buchanan. A simple algorithm for identifying negated findings and diseases in discharge summaries. *J. Biomed. Inform.*, 34:301–310, 2001.

[26] Y.-W. Chen and C.-J. Lin. Combining SVMs with Various Feature Selection Strategies. In I. M. Guyon, S. R. Gunn, M. Nikravesh, and L. Zadeh, editors, *Feature Extraction, Foundations and Applications*. Springer, Berlin, Germany, 2006.

[27] P.-H. Chyou. Patterns of bias due to differential misclassification by casecontrol status in a casecontrol study. *European Journal of Epidemiology*, 22:7–17, 2007.

[28] G. Cipriano, G. Wesenberg, T. Grim, G. N. P. Jr., and M. Gleicher. GRAPE: GRaphical Abstracted Protein Explorer. *Nucleic Acids Research*, 38:W595–W601, 2010.

[29] T. A. Cleary. Test bias: Prediction of grades of negro and white students in integrated colleges. *Journal of Educational Measurement*, 5(2):115–124, 1968.

[30] M. L. Commons, E. J. Trudeau, S. A. Stein, F. A. Richards, and S. R. Krause. Hierarchical complexity of tasks shows the existence of developmental stages. *Developmental Review*, 18:238–278, 1998.

[31] N. Cristianini and J. Shawe-Taylor. *An introduction to Support Vector Machines*. Cambridge University Press, Cambridge, UK, 2002.

[32] J. Davis, E. S. Burnside, I. de Castro Dutra, D. Page, and V. Santos Costa. An integrated approach to learning Bayesian Networks of rules. In *Proceedings of the 16th European Conference on Machine Learning*, pages 84–95, Porto, Portugal, 2005.

[33] J. Davis, E. S. Burnside, I. Dutra, D. Page, R. Ramakrishnan, V. Santos Costa, and J. Shavlik. View Learning for Statistical Relational Learning: With an application to mammography. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence*, pages 677–683, Edinburgh, Scotland, 2005.

[34] J. Davis and M. Goadrich. The relationship between precision-recall and ROC curves. In *Proc. of the 23rd International Conference on Machine Learning*, pages 233–240, Pittsburgh, PA, 2006.

[35] J. Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30, 2006.

[36] R. Díaz-Uriarte and S. A. de Andrés. Gene selection and classification of microarray data using Random Forest. *BMC Bioinformatics*, 7:3, 2006.

[37] B. H. Do, A. Wu, S. Biswal, A. Kamaya, and D. L. Rubin. RADTF: A semantic searchenabled, natural language processorgenerated radiology teaching file. *Radiographics*, 30(7):2039–2048, 2010.

[38] P. C. Dubsky, M. F. X. Gnant, S. Taucher, S. Roka, D. Kandioler, B. Pichler-Gebhard, I. Agstner, M. Seifert, P. Sevelda, and R. Jakesz. Young age as an independent adverse prognostic factor in premenopausal patients with breast cancer. *Clin. Breast Cancer*, 3:65–72, 2002.

[39] S. L. Duggleby, A. A. Jackson, K. M. Godfrey, S. M. Robinson, H. M. Inskip, and the Southampton Womens Survey Study Group. Cut-off points for anthropometric indices of adiposity: differential classification in a large population of young women. *British Journal of Nutrition*, 101:424–430, 2009.

[40] V. L. Ernster, R. Ballard-Barbash, W. E. Barlow, Y. Zheng, D. L. Weaver, G. Cutter, B. C. Yankaskas, R. Rosenberg, P. A. Carney, K. Kerlikowske, S. H. Taplin, N. Urban, and B. M. Geller. Detection of ductal carcinoma in situ in women undergoing screening mammography. *J Natl Cancer Inst*, 94(20):1546–1554, 2002.

[41] E. Eskin. Detecting errors within a corpus using anomaly detection. In *Proc. of the 1st North American chapter of the Association for Computational Linguistics Conference*, pages 148–153, San Francisco, CA, 2000.

[42] K. Fisher. A theory of cognitive development: The control and construction of hierarchies of skills. *Psychological Review*, 87(6):477–531, 1980.

[43] K. M. Flegal, P. M. Keyl, and F. J. Nieto. Differential misclassification arising from nondifferential errors in exposure measurement. *American Journal of Epidemiology*, 134(10):1233–1244, 1991.

[44] M. A. Fox and J. K. Whitesell. *Organic Chemistry*. Jones & Bartlett Publishers, Boston, MA, 3rd edition, 2004.

[45] C. Friedman, P. Alderson, J. Austin, J. Cimino, and S. Johnson. A general natural-language text processor for clinical radiology. *J. Am. Med. Inform. Assn.*, 1(2):161–174, 1994.

[46] C. Friedman, L. Shagina, Y. Lussier, and G. Hripcsak. Automated encoding of clinical documents based on natural language processing. *J. Am. Med. Inform. Assn.*, 11(5):392–402, 2004.

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

[48] C. Gajdos, P. I. Tartter, I. J. Bleiweiss, C. Bodian, and S. T. Brower. Stage 0 to stage III breast cancer in young women. *J. Am. Coll. Surg.*, 190(5):523–529, 2000.

[49] M. Garcia, A. Jemal, E. M. Ward, M. M. Center, Y. Hao, R. L. Siegel, and M. J. Thun. *Global Cancer Facts & Figures 2007*. American Cancer Society, Atlanta, USA, 2007.

[50] E. García-Hernández, R. A. Zubillaga, E. A. Chavelas-Adame, E. Vázquez-Contreras, A. Rojo-Domínguez, and M. Costas. Structural energetics of protein-carbohydrate interactions: Insights derived from the study of lysozyme binding to its natural saccharide inhibitors. *Protein Science*, 12(1):135–142, 2003.

[51] S. Gindl, K. Kaiser, and S. Miksch. Syntactical negation detection in clinical practice guidelines. In *Proc. of the 21st International Congress of the European Federation for Medical Informatics*, pages 187–192, Göteborg, Sweden, 2008.

[52] N. D. Gold and R. M. Jackson. Fold independent structural comparisons of protein-ligand binding sites for exploring functional relationships. *Journal of Molecular Biology*, 355(5):1112–1124, 2006.

[53] C. Goutte and E. Gaussier. A probabilistic interpretation of precision, recall and F-score, with implication for evaluation. In *Proc. of the 27th European Conference on IR Research*, pages 345–359, Santiago de Compostela, Spain, 2005.

[54] N. Guex and M. C. Peitsch. SWISS-MODEL and the Swiss-PdbViewer: An environment for comparative protein modeling. *Electrophoresis*, 18(15):2714–2723, 1997.

[55] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The WEKA Data Mining Software: An update. *SIGKDD Explor. Newsl.*, 11(1):10–18, 2009.

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

[57] H. R. Horton, L. A. Moran, R. S. Ochs, J. D. Rawn, and K. G. Scrimgeour. *Principles of Biochemistry*. Prentice Hall, prentice-hall/pearson education edition, 2002.

[58] Y. Huang and H. Lowe. A novel hybrid approach to automated negation detection in clinical radiology reports. *J. Am. Med. Inform. Assn.*, 14(3):304–311, 2007.

[59] A. K. Jain and B. Chandrasekaran. Dimensionality and Sample Size Considerations in Pattern Recognition Practice. In P. R. Krishnaiah and L. N. Kanal, editors, *Handbook of Statistics*, volume 2, pages 835–855. North-Holland, Amsterdam, 1982.

[60] A. K. Jain, R. P. W. Duin, and J. Mao. Statistical Pattern Recognition: A Review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(1):4–37, January 2000.

[61] R. Kadirvelraj, B. L. Foley, J. D. Dyekjær, and R. J. Woods. Involvement of water in carbohydrate-protein binding: Concanavalin A revisited. *Journal of the American Chemical Society*, 130(50):16933–16942, 2008.

[62] T. Kawabata. Detection of multi-scale pockets on protein surfaces using mathematical morphology. *Proteins*, 78(5):1195–1211, 2010.

[63] L. E. Kelemen, V. S. Pankratz, T. A. Sellers, K. R. Brandt, A. Wang, C. Janney, Z. S. Fredericksen, J. R. Cerhan, and C. M. Vachon. Age-specific trends in mammographic density. *American Journal of Epidemiology*, 167(9):1027–1036, 2008.

[64] J. Kelsey, A. Whittemore, W. Thompson, and E. A. *Methods in observational epidemiology*. Oxford University Press, USA, 1996.

[65] S. Khuri, F. T. Bakker, and J. M. Dunwell. Phylogeny, function and evolution of the cupins, a structurally conserved, functionally diverse superfamily of proteins. *Molecular Biology and Evolution*, 18(4):593–605, 2001.

[66] S. H. Kim, B. K. Seo, J. Lee, S. J. Kim, K. R. Cho, K. Y. Lee, B.-K. Je, H. Y. Kim, Y.-S. Kim, and J.-H. Lee. Correlation of ultrasound findings with histology, tumor grade, and biological markers in breast cancer. *Acta Oncol.*, 47(8):1531–1538, 2008.

[67] I. Kononenko, E. Simec, and M. Robnik-Sikonja. Overcoming the myopia of inductive learning algorithms with RELIEFF. *Appl. Intell.*, 7(1):39–55, 1997.

[68] D. B. Kopans, R. H. Moore, K. A. McCarthy, D. A. Hall, C. A. Hulka, G. J. Whitman, P. J. Slanetz, and E. F. Halpern. Positive predictive value of breast biopsy performed as a result of mammography: there is no abrupt change at age 50 years. *Radiology*, 200(2):357 – 360, August 1996.

[69] C. P. Lam and D. G. Stork. Evaluating classifiers by means of test data with noisy labels. In *Proc. of the 18th International Joint Conference on Artificial Intelligence*, pages 513–518, Acapulco, Mexico, 2003.

[70] J. Larson and R. S. Michalski. Inductive inference of VL decision rules. *ACM SIGART Bulletin*, 63:38–44, June 1977.

[71] D. A. Lindberg, B. L. Humphreys, and A. T. McCray. The unified medical language system. *Method. Inform. Med.*, 32:281–291, 1993.

[72] R. L. Linn. Single-group validity, differential validity, and differential prediction. *Journal of Applied Psychology*, 63:507–512, 1978.

[73] R. L. Linn and C. E. Werts. Considerations for studies of test bias. *Journal of Educational Measurement*, 8:1–4, 1971.

[74] Y. Liu, M. Pérez, M. Schootman, R. L. Aft, W. E. Gillanders, M. J. Ellis, and D. B. Jeffe. A longitudinal study of factors associated with perceived risk of recurrence in women with ductal carcinoma in situ and early-stage invasive breast cancer. *Breast Cancer Res. Treat.*, Epub ahead of print, 2010.

[75] V. S. Lo. The true lift model - a novel data mining approach to response modeling in database marketing. *SIGKDD Explorations*, 4(2):78–86, 2002.

[76] W. Long. Lessons extracting diseases from discharge summaries. In *American Medical Informatics Association Annual Symposium Proceedings*, pages 478–482, Chicago, IL, 2007.

[77] A. Malik and S. Ahmad. Sequence and structural features of carbohydrate binding in proteins and assessment of predictability using a Neural Network. *BMC Structural Biology*, 7:1, 2007.

[78] M. T. Mandelson, N. Oestreicher, P. L. Porter, D. White, C. A. Finder, S. H. Taplin, and E. White. Breast density as a predictor of mammographic detection: comparison of interval- and screen-detected cancers. *J. Natl. Cancer Inst.*, 92(13):1081–1087, 2000.

[79] T. M. Mitchell. *Machine Learning*. McGraw-Hill International Editions, Singapore, 1997.

[80] S. Muggleton. Random train generator, 1998.

[81] S. Muggleton and C. Feng. Efficient induction of logic programs. In *Proceedings of the 1st Conference on Algorithmic Learning Theory*, pages 368–381, Tokyo, 1990.

[82] S. Muggleton, J. Santos, and A. Tamaddoni-Nezhad. ProGolem: a system based on relative minimal generalisation. In *Proceedings of the 19th International Conference on ILP, Springer*, pages 131–148, Leuven, Belgium, 2009.

[83] S. H. Muggleton. Inverse entailment and Progol. *New Generation Computing*, 13:245–286, 1995.

[84] P. G. Mutalik, A. Deshpande, and P. M. Nadkarni. Use of general-purpose negation detection to augment concept indexing of medical documents: A quantitative study using the UMLS. *J. Am. Med. Inform. Assn.*, 8(6):598–609, 2001.

[85] R. Nakayama, Y. Uchiyama, R. Watanabe, S. Katsuragawa, K. Namba, and K. Doi. Computer-aided diagnosis scheme for histological classification of clustered microcalcifications on magnification mammograms. *Med. Phys.*, 31(4):789–799, 2004.

[86] H. Nassif, H. Al-Ali, S. Khuri, and W. Keirouz. Prediction of protein-glucose binding sites using Support Vector Machines. *Proteins: Structure, Function, and Bioinformatics*, 77(1):121–132, 2009.

[87] H. Nassif, H. Al-Ali, S. Khuri, W. Keirouz, and D. Page. An Inductive Logic Programming approach to validate hexose biochemical knowledge. In *Proceedings of the 19th International Conference on ILP*, pages 149–165, Leuven, Belgium, 2009.

[88] H. Nassif, F. Cunha, I. Moreira, R. Cruz-Correia, E. Sousa, D. Page, E. Burnside, and I. Dutra. Extracting BI-RADS features from Portuguese clinical texts. In *BIBM'12*, Philadelphia, USA, 2012. Accepted.

[89] H. Nassif, D. Page, M. Ayvaci, J. Shavlik, and E. S. Burnside. Uncovering age-specific invasive and DCIS breast cancer rules using Inductive Logic Programming. In *1st ACM International Health Informatics Symposium*, pages 76–82, Arlington, VA, 2010.

[90] H. Nassif, V. Santos Costa, E. S. Burnside, and D. Page. Relational differential prediction. In *ECML'12*, pages 617–632, Bristol, UK, 2012.

[91] H. Nassif, R. Wood, E. S. Burnside, M. Ayvaci, J. Shavlik, and D. Page. Information extraction for clinical data mining: A mammography case study. In *ICDM Workshops*, pages 37–42, Miami, Florida, 2009.

[92] H. Nassif, Y. Wu, D. Page, and E. S. Burnside. Logical Differential Prediction Bayes Net, improving breast cancer diagnosis for older women. In *AMIA'12*, Chicago, 2012. Accepted.

[93] H. B. Nichols, A. Trentham-Dietz, J. M. Hampton, L. Titus-Ernstoff, K. M. Egan, W. C. Willett, and P. A. Newcomb. From menarche to menopause: Trends among us women born from 1912 to 1969. *American Journal of Epidemiology*, 164(10):1003–1011, 2006.

[94] K. Nigam, A. McCallum, S. Thrun, and T. Mitchell. Learning to classify text from labeled and unlabeled documents. In *Proc. of the 15th National Conference on Artificial Intelligence*, pages 792–799, 1998.

[95] P. V. Ogren. Knowtator: a protégé plug-in for annotated corpus construction. In *Proceedings of the 2006 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 273–275, 2006.

[96] D. Page, V. Santos Costa, S. Natarajan, A. Barnard, P. Peissig, and M. Caldwell. Identifying adverse drug events by relational learning. In *AAAI-12*, pages 1599–1605, Toronto, 2012.

[97] N. Patani, B. Cutuli, and K. Mokbel. Current management of DCIS: a review. *Breast Cancer Res Treat*, 111(1):1–10, 2008.

[98] M. Patra and C. Mandal. Search for glucose/galactose-binding proteins in newly discovered protein sequences using molecular modeling techniques and structural analysis. *Glycobiology*, 16(10):959–968, 2006.

[99] B. Percha, H. Nassif, J. Lipson, E. Burnside, and D. Rubin. Automatic classification of mammography reports by BI-RADS breast tissue composition class. *J. Am. Med. Inform. Assn.*, 19(5):913–916, 2012.

[100] M. Petticrew, A. Sowden, and D. Lister-Sharp. False-negative results in screening programs: Medical, psychological, and other implications. *Int. J. Technol. Assess.*, 17(2):164–170, 2001.

[101] B. Phibbs and W. Nelson. Differential classification of acute myocardial infarction into ST- and Non-ST segment elevation is not valid or rational. *Annals of Noninvasive Electrocardiology*, 15(3):191–199, 2010.

[102] F. A. Quiocho and N. K. Vyas. Atomic interactions between proteins/enzymes and carbohydrates. In S. M. Hecht, editor, *Bioorganic Chemistry: Carbohydrates*, chapter 11, pages 441–457. Oxford University Press, New York, 1999.

[103] N. J. Radcliffe and P. D. Surry. Differential response analysis: Modeling true response by isolating the effect of a single action. In *Credit Scoring and Credit Control VI*, Edinburgh, Scotland, 1999.

[104] N. J. Radcliffe and P. D. Surry. Real-world uplift modelling with significance-based uplift trees. White Paper TR-2011-1, Stochastic Solutions, 2011.

[105] V. S. R. Rao, K. Lam, and P. K. Qasba. Architecture of the sugar binding sites in carbohydrate binding proteins—a computer modeling study. *International Journal of Biological Macromolecules*, 23(4):295–307, 1998.

[106] L. Rokach, O. Maimon, and M. Averbuch. Information retrieval system for medical narrative reports. In *Proc. of the 6th International Conference on Flexible Query Answering Systems*, pages 217–228, Lyon, France, 2004.

[107] R. Romano, L. Rokach, and O. Maimon. Cascaded data mining methods for text understanding, with medical case study. In *Proc. of the 6th IEEE International Conference on Data Mining - Workshops*, Hong Kong, China, 2006.

[108] R. D. Rosenberg, W. C. Hunt, M. R. Williamson, F. D. Gilliland, P. W. Wiest, C. A. Kelsey, C. R. Key, and M. N. Linver. Effects of age, breast density, ethnicity, and estrogen replacement therapy on screening mammographic sensitivity and cancer stage at diagnosis: review of 183,134 screening mammograms in Albuquerque, New Mexico. *Radiology*, 209(2):511–518, 1998.

[109] P. Ruch, R. Baud, A. Geissbuhler, and A. M. Rassinoux. Comparing general and medical texts for information retrieval based on natural language processing: an inquiry into lexical disambiguation. In *Proc. of the 10th World Congress on Medical Informatics*, volume 10 (Pt 1), pages 261–265, London, UK, 2001.

[110] S. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach*. Prentice Hall, 3rd edition, 2009.

[111] P. Rzepakowski and S. Jaroszewicz. Decision trees for uplift modeling. In *2010 IEEE International Conference on Data Mining*, pages 441–450, Sydney, Australia, 2010.

[112] P. R. Sackett, R. M. Laczo, and Z. P. Lippe. Differential prediction and the use of multiple predictors: The omitted variables problem. *Journal of Applied Psychology*, 88(6):1046–1056, 2003.

[113] M. Sahami, S. Dumais, D. Heckerman, and E. Horvitz. A bayesian approach to filtering junk e-mail. In *AAAI Workshop on Learning for Text Categorization*, Madison, WI, 1998.

[114] J. C. A. Santos, H. Nassif, D. Page, S. H. Muggleton, and M. J. E. Sternberg. Automated identification of protein-ligand interaction features using Inductive Logic Programming: A hexose binding case study. *BMC Bioinformatics*, 13:162, 2012.

[115] V. Santos Costa. The life of a logic programming system. In M. G. de la Banda and E. Pontelli, editors, *Proceedings of the 24th International Conference on Logic Programming*, pages 1–6, Udine, Italy, 2008.

[116] J. Screen, E. C. Stanca-Kaposta, D. P. Gamblin, B. Liu, N. A. Macleod, L. C. Snoek, B. G. Davis, and J. P. Simons. IR-spectral signatures of aromaticsugar complexes: Probing carbohydrateprotein interactions. *Angew. Chem. Int. Ed.*, 46:3644–3648, 2007.

[117] C. Shionyu-Mitsuyama, T. Shirai, H. Ishida, and T. Yamane. An empirical approach for structure-based prediction of carbohydrate-binding sites on proteins. *Protein Engineering*, 16(7):467–478, 2003.

[118] P. Smyth. Bounds on the mean classification error rate of multiple experts. *Pattern Recogn. Lett.*, 17(12):1253–1257, 1996.

[119] Society for Industrial and Organizational Psychology. *Principles for the Validation and Use of Personnel Selection Procedures*, 4th edition, 2003.

[120] E. Solomon, L. Berg, and D. W. Martin. *Biology*. Brooks Cole, Belmont, CA, 8th edition, 2007.

[121] A. Srinivasan. *The Aleph Manual*, 4th edition, 2007.

[122] Z. Stein and K. Heikkinen. Models, metrics, and measurement in developmental psychology. *Integral Review*, 5(1):4–24, 2009.

[123] M. S. Sujatha and P. V. Balaji. Identification of common structural features of binding sites in galactose-specific proteins. *Proteins*, 55(1):44–65, 2004.

[124] M. S. Sujatha, Y. U. Sasidhar, and P. V. Balaji. Energetics of galactose- and glucose-aromatic amino acid interactions: Implications for binding in galactose-specific proteins. *Protein Science*, 13(9):2502–2514, 2004.

[125] S. A. Sullivan and D. Landsman. Characterization of sequence variability in nucleosome core histone folds. *Proteins: Structure, Function, and Genetics*, 52:454–465, 2003.

[126] L. Tabar, H. H. Tony Chen, M. F. Amy Yen, T. Tot, T. H. Tung, L. S. Chen, Y. H. Chiu, S. W. Duffy, and R. A. Smith. Mammographic tumor features can predict long-term outcomes reliably in women with 1-14-mm invasive breast carcinoma. *Cancer*, 101(8):1745–1759, 2004.

[127] C. Taroni, S. Jones, and J. M. Thornton. Analysis and prediction of carbohydrate binding sites. *Protein Eng.*, 13(2):89–98, 2000.

[128] M. G. Thurfjell, A. Lindgren, and E. Thurfjell. Nonpalpable breast cancer: Mammographic appearance as predictor of histologic type. *Radiology*, 222(1):165–170, 2002.

[129] V. N. Vapnik. *Statistical Learning Theory*. John Wiley & Sons, New York, 1998.

[130] F. A. Vicini and A. Recht. 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.

[131] A. M. M. Vlaar, A. Bouwmans, W. H. Mess, S. C. Tromp, and W. E. J. Weber. Transcranial duplex in the differential diagnosis of parkinsonian syndromes: a systematic review. *Journal of Neurology*, 256(4):530–538, 2009.

[132] A. Vlachos and M. Craven. Detecting speculative language using syntactic dependencies and logistic regression. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, pages 18–25, Uppsala, Sweden, July 2010.

[133] L. Wall and R. L. Schwartz. *Programming Perl*. O'Reilly & Associates, Sebastopol, CA, United States of America, 1992.

[134] G. Wang and R. L. Dunbrack. PISCES: A Protein Sequence Culling Server. *Bioinformatics*, 19(12):1589–1591, 2003.

[135] J. N. Wolfe. Breast parenchymal patterns and their changes with age. *Radiology*, 121:545–552, 1976.

[136] G. Y. Wong and F. H. Leung. Predicting protein-ligand binding site with support vector machine. In *Proceedings of the IEEE Congress on Evolutionary Computation*, pages 1–5, 2010.

[137] J. W. Young. Differential validity, differential prediction, and college admissions testing: A comprehensive review and analysis. Research Report 2001-6, The College Board, New York, 2001.

[138] F. Zelezný and N. Lavrac. Propositionalization-based relational subgroup discovery with RSD. *Machine Learning*, 62(1-2):33–66, 2006.

[139] Y. Zhang, G. J. Swaminathan, A. Deshpande, E. Boix, R. Natesh, Z. Xie, K. R. Acharya, and K. Brew. Roles of individual enzyme-substrate interactions by alpha-1,3-galactosyltransferase in catalysis and specificity. *Biochemistry*, 42(46):13512–13521, 2003.

Appendix A: Hexose Dataset

Table A.1: Inventory of the hexose-binding positive data set

Hexose	PDB ID	Ligand	PDB ID	Ligand	PDB ID	Ligand
Glucose	1BDG	GLC-501	1ISY	GLC-1471	1SZ2	BGC-1001
	1EX1	GLC-617	1J0Y	GLC-1601	1SZ2	BGC-2001
	1GJW	GLC-701	1JG9	GLC-2000	1U2S	GLC-1
	1GWW	GLC-1371	1K1W	GLC-653	1UA4	GLC-1457
	1H5U	GLC-998	1KME	GLC-501	1V2B	AGC-1203
	1HIZ	GLC-1381	1MMU	GLC-1	1WOQ	GLC-290
	1HIZ	GLC-1382	1NF5	GLC-125	1Z8D	GLC-901
	1HKC	GLC-915	1NSZ	GLC-1400	2BQP	GLC-337
	1HSJ	GLC-671	1PWB	GLC-405	2BVW	GLC-602
	1HSJ	GLC-672	1Q33	GLC-400	2BVW	GLC-603
	1I8A	GLC-189	1RYD	GLC-601	2F2E	AGC-401
	1ISY	GLC-1461	1S5M	AGC-1001		
Galactose	1AXZ	GLA-401	1MUQ	GAL-301	1R47	GAL-1101
	1DIW	GAL-1400	1NS0	GAL-1400	1S5D	GAL-704
	1DJR	GAL-1104	1NS2	GAL-1400	1S5E	GAL-751
	1DZQ	GAL-502	1NS8	GAL-1400	1S5F	GAL-104
	1EUU	GAL-2	1NSM	GAL-1400	1SO0	GAL-500
	1ISZ	GAL-461	1NSU	GAL-1400	1TLG	GAL-1
	1ISZ	GAL-471	1NSX	GAL-1400	1UAS	GAL-1501
	1JZ7	GAL-2001	1OKO	GLB-901	1UGW	GAL-200
	1KWK	GAL-701	1OQL	GAL-265	1XC6	GAL-9011
	1L7K	GAL-500	1OQL	GAL-267	1ZHJ	GAL-1
	1LTI	GAL-104	1PIE	GAL-1	2GAL	GAL-998
Mannose	1BQP	MAN-402	1KZB	MAN-1501	1OUR	MAN-301
	1KLF	MAN-1500	1KZC	MAN-1001	1QMO	MAN-302
	1KX1	MAN-20	1KZE	MAN-1001	1U4J	MAN-1008
	1KZA	MAN-1001	1OP3	MAN-503	1U4J	MAN-1009

Table A.2: Inventory of the non-hexose-binding negative data set

PDB ID	Cavity Center	Ligand	PDB ID	Cavity Center	Ligand
Hexose-like ligands					
1A8U	4320, 4323	BEZ-1	1AI7	6074, 6077	IPH-1
1AWB	4175, 4178	IPD-2	1DBN	pyranose ring	GAL-102
1EOB	3532, 3536	DHB-999	1F9G	5792, 5785, 5786	ASC-950
1G0H	4045, 4048	IPD-292	1JU4	4356, 4359	BEZ-1
1LBX	3941, 3944	IPD-295	1LBY	3944, 3939, 3941	F6P-295
1LIU	15441, 15436, 15438	FBP-580	1MOR	pyranose ring	G6P-609
1NCW	3406, 3409	BEZ-601	1P5D	pyranose ring	G1P-658
1T10	4366, 4361, 4363	F6P-1001	1U0F	pyranose ring	G6P-900
1UKB	2144, 2147	BEZ-1300	1X9I	pyranose ring	G6Q-600
1Y9G	4124, 4116, 4117	FRU-801	2B0C	pyranose ring	G1P-496
2B32	3941, 3944	IPH-401	4PBG	pyranose ring	BGP-469
Other ligands					
11AS	5132	ASN-1	11GS	1672, 1675	MES-3
1A0J	6985	BEN-246	1A42	2054, 2055	BZO-555
1A50	4939, 4940	FIP-270	1A53	2016, 2017	IGP-300
1AA1	4472, 4474	3PG-477	1AJN	6074, 6079	AAN-1
1AJS	3276, 3281	PLA-415	1AL8	2652	FMN-360
1B8A	7224	ATP-500	1BO5	7811	GOL-601
1BOB	2566	ACO-400	1D09	7246	PAL-1311
1EQY	3831	ATP-380	1IOL	2674, 2675	EST-400
1JTV	2136, 2137	TES-500	1KF6	16674, 16675	OAA-702
1RTK	3787, 3784	GBS-300	1TJ4	1947	SUC-1
1TVO	2857	FRZ-1001	1UK6	2142	PPI-1300
1W8N	4573, 4585	DAN-1649	1ZYU	1284, 1286	SKM-401
2D7S	3787	GLU-1008	2GAM	11955	NGA-502
3PCB	3421, 3424	3HB-550			

Table A.3: Inventory of the non-binding surface groove negative data set

PDB ID	Cavity Center	PDB ID	Cavity Center	PDB ID	Cavity Center
1A04	1424, 2671	1A0I	1689, 799	1A22	2927
1AA7	579	1AF7	631, 1492	1AM2	1277
1ARO	154, 1663	1ATG	1751	1C3G	630, 888
1C3P	1089, 1576	1DXJ	867, 1498	1EVT	2149, 2229
1FI2	1493	1KLM	4373, 4113	1KWP	1212
1QZ7	3592, 2509	1YQZ	4458, 4269	1YVB	1546, 1814
1ZT9	1056, 1188	2A1K	2758, 3345	2AUP	2246
2BG9	14076, 8076	2C9Q	777	2CL3	123, 948
2DN2	749, 1006	2F1K	316, 642	2G50	26265, 31672
2G69	248, 378	2GRK	369, 380	2GSE	337, 10618
2GSH	6260				

Appendix B: Logistic Regression Models

Table B.1: Older cohort multivariable model using stepwise regression with AIC criterion

Risk Factor	Value	Beta	Odds Ratio	95% CI	p-value
	(Intercept)	-1.16	0.31	0.18–0.55	0.000 ***
Palpable Lump					0.013 **
	No Corresponding Palpable Mass	0.00	1(referent)		
	Missing	-0.30	0.74	0.05–10.55	0.824
	Corresponding Palpable Mass	0.80	2.22	1.12–4.41	0.022 **
Family History					0.043 **
	None	0.00	1(referent)		
	Missing	-0.89	0.41	0.13–1.32	0.135
	Strong	-0.32	0.73	0.33–1.59	0.422
	Very Strong	1.66	5.24	0.84–32.78	0.076 *
Prior Surgery					0.132
	Not Present	0.00	1(referent)		
	Missing	-0.36	0.70	0.07–6.82	0.759
	Present	0.57	1.78	0.99–3.17	0.053 *
Principal Abnormal Finding					0.000 ***
	Calcifications or Single Dilated Duct	0.00	1(referent)		
	Architectural Distortion	20.56	Inf	0.00–Inf	0.993
	Associated Calcifications	2.16	8.67	3.39–22.14	0.000 ***
	Missing	2.10	8.14	3.88–17.09	0.000 ***

continued on the next page

Table B.1 – *continued from previous page*

Risk Factor	Value	Beta	Odds Ratio	95% CI	p-value
	Asymmetry or Focal Asymmetry	2.94	18.87	3.79–93.87	0.000 ***
	Mass	3.04	20.93	9.20–47.65	0.000 ***
	Developing Asymmetry	2.80	16.45	1.78–151.95	0.014 **
Calcification Distribution					0.008 **
	Not Present	0.00	1(referent)		
	Linear or Segmental	-3.11	0.04	0.00–0.49	0.011 **
	Clustered	-0.69	0.50	0.22–1.18	0.113
	Regional or Scattered	-1.94	0.14	0.01–2.83	0.202
Mass Margins					0.000 ***
	None	0.00	1(referent)		
	Circumscribed	-2.51	0.08	0.01–0.45	0.004 ***
	Ill-Defined	0.19	1.21	0.46–3.20	0.703
	Obscured	0.10	1.10	0.11–11.31	0.935
	Spiculated	28.70	Inf	0.00–Inf	0.983
Mass Shape					0.033 **
	None	0.00	1(referent)		
	Irregular	1.91	6.78	0.78–58.79	0.083 *
	Lobular or Oval	-0.13	0.87	0.24–3.16	0.838
	Round	-15.53	0.00	0.00–Inf	0.987
Focal Asymmetric Density					0.077 *
	Not Present	0.00	1(referent)		
	Present	1.63	5.10	0.54–47.77	0.154

Levels of significance: *** p-value < 0.001; ** p-value < 0.05 and * p-value < 0.1

Table B.2: Younger cohort multivariable model using stepwise regression with AIC criterion

Risk Factor	Value	Beta	Odds Ratio	95% CI	p-value
	(Intercept)	-0.64	0.53	0.35–0.80	0.002 ***
Palpable Lump					0.000 ***
	No Corresponding Palpable Mass	0.00	1(referent)		
	Missing	-0.68	0.51	0.16–1.60	0.246
	Corresponding Palpable Mass	1.21	3.36	1.79–6.32	0.000 ***
Principal Abnormal Finding					0.000 ***
	Calcifications or Single Dilated Duct	0.00	1(referent)		
	Architectural Distortion	1.95	7.05	0.75–65.98	0.087 *
	Associated Calcifications	1.58	4.85	1.87–12.55	0.001 ***
	Missing	1.02	2.76	1.34–5.70	0.006 ***
	Asymmetry or Focal Asymmetry	1.86	6.41	1.26–32.64	0.025 **
	Mass	2.74	15.51	4.97–48.35	0.000 ***
	Developing Asymmetry	16.50	Inf	0.00–Inf	0.997
Architectural Distortion					0.063
	Not Present	0.00	1(referent)		
	Present	1.78	5.91	0.67–52.13	0.110
Mass Shape					0.090 *
	None	0.00	1(referent)		
	Irregular	15.83	Inf	0.00–Inf	0.986
	Lobular or Oval	0.09	1.10	0.23–5.21	0.787

continued on the next page

Table B.2 – *continued from previous page*

Risk Factor	Value	Beta	Odds Ratio	95% CI	p-value
	Round	-19.53	0.00	0.00–Inf	0.996
Mass Size (mm)					0.047 *
	None	0.00	1(referent)		
	10–20	-0.97	0.38	0.03–4.61	0.447
	20–50	1.70	5.47	1.17–25.69	0.031 **
	< 10	-0.58	0.56	0.19–1.63	0.287
	>= 50	-0.58	0.56	0.14–2.25	0.413

Levels of significance: *** p-value < 0.001; ** p-value < 0.05 and * p-value < 0.1