Differential Relational Learning

Houssam Nassif

Thesis Defense
03 August 2012
Outline

1. Differential Prediction
2. Expert Driven Approach
   - Expert Driven Method
   - ProGolem Recall *
3. Model Filtering Approach
4. Differential Prediction Search Approach
5. BI-RADS Information Extraction *
6. Other Work *
Outline

1. Differential Prediction

2. Expert Driven Approach
   - Expert Driven Method
   - ProGolem Recall *

3. Model Filtering Approach

4. Differential Prediction Search Approach

5. BI-RADS Information Extraction *

6. Other Work *
Breast-Cancer Stages

In-Situ Cancer Stage
Breast-Cancer Stages

Invasive Cancer Stage

Basement Membrane

Abnormal cells
Cancer Stage Features

- In Situ can develop into Invasive
  - Current practice: Always treat In Situ
- Time to spread may be very long
  - Patient may die of other causes
  - Over-diagnosis (and over-treatment)

- What features characterize In Situ in older patients?
- What features change between older and younger?
In situ cancer can develop into invasive cancer.

- Current practice: Always treat in situ.

- Time to spread may be very long.
  - Patient may die of other causes.
  - Over-diagnosis (and over-treatment).

What features characterize in situ cancer in older patients?

What features change between older and younger patients?
Definition (Cleary’68)

Differential Prediction (DP): case where consistent nonzero errors of prediction are made for members of a given subgroup.
Definition (Cleary’68)

Differential Prediction (DP): case where consistent nonzero errors of prediction are made for members of a given subgroup.
Definition (*Cleary’68*)

Differential Prediction (DP): case where consistent nonzero errors of prediction are made for members of a given subgroup.
Differential Prediction (Definition (Cleary’68))

Differential Prediction (DP): case where consistent nonzero errors of prediction are made for members of a given subgroup.
Using Regression to Detect DP

- Validate educational and psychological tests
- Detect discrepancies related to race or gender
Using Regression to Detect DP

- Validate educational and psychological tests
- Detect discrepancies related to race or gender

Graphs showing the relationship between SAT scores and 1st Year GPA for males (M) and females (F).
Using Regression to Detect DP

- Validate educational and psychological tests
- Detect discrepancies related to race or gender

Diagram showing relationship between SAT scores and 1st Year GPA for males (M) and females (F).
DP in Machine Learning

- Detected by:
  - Comparing classifiers built on distinct data subgroups
  - Checking classifier performance on multiple subgroups

- Uplift Modeling in marketing

- Limited to non-relational datasets

- Related to:
  - Relational Subgroup Discovery (Zelezný’06)
  - Differential misclassification cost
Aim

- Classifier to maximize DP over given data subsets
- Extend DP to relational sets
- Insight into DP features

Thesis Statement

ILP-based DP can:

- Propose rules over multi-relational datasets
- Maximize differences over given subsets (strata)
- Offer insight into underlying domain
Stratified Dataset

Strata are disjoint, each strata should contain at least one example of each target class

Definition (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:

$$(\forall (i, j) \in [1, n], i \neq j) \quad D_i \subset D, \quad D_i \cap D_j = \emptyset, \quad \forall l \quad D_i^l \neq \emptyset. \quad (1)$$
Differential Predictive Rule

Given a stratified data, a rule whose performance is significantly better over one stratum as compared to the others.

Definition (Differential Predictive Rule)

Let $c$ be a rule over the set of instances $X$, and let $D$ be a stratified dataset. Let $S(c|D_i)$ be the classification performance score for $c$ over the subset $D_i$. A stratum-$j$ specific differential predictive rule is a rule $c_j$ such that:

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

Score difference ($\gg$) can be evaluated using statistical significance tests, a tuning set, or a threshold.
Inductive Logic Programming

**Definition**

Inductive Logic Programming (ILP): Machine learning approach that learns a set of first-order logic rules that explain the data

1. Generates easy to interpret if-then rules
2. Allows user interaction through background knowledge
3. Operates on relational datasets
4. Can investigate the performance of each rule, selecting for DP over given subsets
Inductive Logic Programming

Definition

Inductive Logic Programming (ILP): Machine learning approach that learns a set of first-order logic rules that explain the data.

1. Generates easy to interpret if-then rules
2. Allows user interaction through background knowledge
3. Operates on relational datasets
4. Can investigate the performance of each rule, selecting for DP over given subsets
Outline

1. Differential Prediction

2. Expert Driven Approach
   - Expert Driven Method
   - ProGolem Recall *

3. Model Filtering Approach

4. Differential Prediction Search Approach

5. BI-RADS Information Extraction *

6. Other Work *
Expert Driven Approach

- Simplest method
- Build a classifier on each strata
- Expert compares both classifiers and infers DP rules

- Expert can be a human or a machine
- Method can be applied to non-rule-learning classifiers
Expert Driven Approach

- Simplest method
- Build a classifier on each strata
- Expert compares both classifiers and infers DP rules
- Expert can be a human or a machine
- Method can be applied to non-rule-learning classifiers
Hexose-Binding Modeling

- Galactose, glucose, mannose
- High specificity to diverse protein families
- Interesting to uncover differential binding patterns between glucose-binding and general hexose-binding
Hexose-Binding Modeling

- Galactose, glucose, mannose
- High specificity to diverse protein families
- Interesting to uncover differential binding patterns between glucose-binding and general hexose-binding
Hexose Binding-Site Representation
Hexose Binding-Site Features

1: **procedure** EXTRACTFEATURES(binding site center)
2:      **for all** concentric layers **do**
3:          **for all** PDB atoms **do**
4:              get distance from center
5:              get charge
6:              get hydrophobicity
7:              get hydrogen-bonding
8:              get residue
9:          **end for**
10:      **end for**
11: **end procedure**
Building Classifiers

Hexose and glucose classifiers:

- SVM + RF
  - Random Forests for feature selection
  - Support Vector Machines for classification
- ILP (Aleph)

- First glucose model/classifier
- Data-driven validation of hexose-binding biological knowledge
Building Classifiers

Hexose and glucose classifiers:

- SVM + RF
  - Random Forests for feature selection
  - Support Vector Machines for classification

- ILP (Aleph)

- First glucose model/classifier

- Data-driven validation of hexose-binding biological knowledge
## Expert Driven Differential Rules

<table>
<thead>
<tr>
<th>Feature</th>
<th>Glucose</th>
<th>Hexose</th>
</tr>
</thead>
<tbody>
<tr>
<td>Water and ions</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Negative charge and carboxylate residue</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Surface hydrogen bonding</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Dual hydrophobic-hydrophilic</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Aromatic residue docking</td>
<td></td>
<td>x</td>
</tr>
</tbody>
</table>

- Hydrophobic stacking: hexose ring over aromatic ring
- Glucose stacks over non-aromatic residues
### Expert Driven Differential Rules

<table>
<thead>
<tr>
<th>Feature</th>
<th>Glucose</th>
<th>Hexose</th>
</tr>
</thead>
<tbody>
<tr>
<td>Water and ions</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Negative charge and carboxylate residue</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Surface hydrogen bonding</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Dual hydrophobic-hydrophilic</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Aromatic residue docking</td>
<td></td>
<td>x</td>
</tr>
</tbody>
</table>

- Hydrophobic stacking: hexose ring over aromatic ring
- Glucose stacks over non-aromatic residues
Outline

1. Differential Prediction
2. Expert Driven Approach
   - Expert Driven Method
   - ProGolem Recall *
3. Model Filtering Approach
4. Differential Prediction Search Approach
5. BI-RADS Information Extraction *
6. Other Work *
**ILP Example**

- Pick a positive instance
  - $P(X)$ if $\text{square}(X)$
  - $P(X)$ if $\text{red}(X) \land \text{big}(X)$
  - 1 false positive
  - $P(X)$ if $\text{sibling}(X, Y) \land \text{square}(Y)$
  - 1 false negative
  - Form theory

**Example**

```latex
P(A), \text{red}(A), \text{big}(A), \text{round}(A)
\text{sibling}(A, B)
```
ILP Example

Example

\[ P(A), \text{red}(A), \text{big}(A), \text{round}(A), \text{sibling}(A, B) \]
ILP Example

Example

\[ P(A), \text{red}(A), \text{big}(A), \text{round}(A), \text{ sibling}(A, B) \]

Pick a positive instance

- \( P(X) \) if square\((X) \)
- \( P(X) \) if \( \text{red}(X) \land \text{big}(X) \)

1 false positive

- \( P(X) \) if \( \text{sibling}(X, Y) \land \text{square}(Y) \)

1 false negative

Form theory
**ILP Example**

Example

\[ P(A), \text{red}(A), \text{big}(A), \text{round}(A), \text{sibling}(A, B) \]

- Pick a positive instance
- \( P(X) \) if \( \text{square}(X) \)
- \( P(X) \) if \( \text{red}(X) \land \text{big}(X) \)
- 1 false positive
- \( P(X) \) if \( \text{sibling}(X, Y) \land \text{square}(Y) \)
- 1 false negative
- Form theory
ILP Example

Example

\( P(A), \text{red}(A), \text{big}(A), \text{round}(A) \), sibling(A, B)

- Pick a positive instance
- \( P(X) \) if \( \text{square}(X) \)
- \( P(X) \) if \( \text{red}(X) \land \text{big}(X) \)
- 1 false positive
- \( P(X) \) if \( \text{sibling}(X, Y) \land \text{square}(Y) \)
- 1 false negative
- Form theory
Example

\[ P(A), \text{red}(A), \text{big}(A), \text{round}(A), \text{siblings}(A, B) \]
ILP Example

Example

\( P(A), \text{red}(A), \text{big}(A), \text{round}(A) \)
\( \text{sibling}(A, B) \)

Pick a positive instance
- \( P(X) \text{ if square}(X) \)
- \( P(X) \text{ if red}(X) \land \text{big}(X) \)
  - 1 false positive
- \( P(X) \text{ if sibling}(X, Y) \land \text{square}(Y) \)
  - 1 false negative

Form theory
ILP Example

Example

\[ P(A), \text{red}(A), \text{big}(A), \text{round}(A) \]
\[ \text{sibling}(A, B) \]
ILP Example

Example

\[ P(A), \text{red}(A), \text{big}(A), \text{round}(A), \text{sibling}(A, B) \]

Pick a positive instance

- \( P(X) \) if \( \text{square}(X) \)
- \( P(X) \) if \( \text{red}(X) \land \text{big}(X) \)
  - 1 false positive
- \( P(X) \) if \( \text{sibling}(X, Y) \land \text{square}(Y) \)
  - 1 false negative
- Form theory
Example

\[ P(A), \text{red}(A), \text{big}(A), \text{round}(A) \]
\[ \text{sibling}(A, B) \]
ILP Search

Pick a positive instance
- Construct the Bottom Clause, most specific clause
- Top-down search: Start with most general rule, add bottom clause predicates (Aleph, Srinivasan’07)
- Bottom-up search: Start with bottom clause, remove predicates (ProGolem, Muggleton’09)

Example (Bottom Clause (A))

red(A), big(A), round(A), sibling(A, B),
red(B), big(B), round(B)
ILP Search

- Pick a positive instance
- Construct the *Bottom Clause*, most specific clause

*Top-down search:* Start with most general rule, add bottom clause predicates (Aleph, Srinivasan’07)

*Bottom-up search:* Start with bottom clause, remove predicates (ProGolem, Muggleton’09)

Example (Bottom Clause (A))

\[
\text{red}(A), \text{big}(A), \text{round}(A), \\
\text{Sibling}(A, B), \\
\text{red}(B), \text{big}(B), \text{round}(B)
\]
ILP Search

Example (Bottom Clause (A))

red(A), big(A), round(A), sibling(A, B), red(B), big(B), round(B)

- Pick a positive instance
- Construct the *Bottom Clause*, most specific clause
- **Top-down search**: Start with most general rule, add bottom clause predicates (Aleph, Srinivasan’07)
- **Bottom-up search**: Start with bottom clause, remove predicates (ProGolem, Muggleton’09)
ILP Search

- Pick a positive instance
- Construct the *Bottom Clause*, most specific clause
- **Top-down search**: Start with most general rule, add bottom clause predicates (Aleph, Srinivasan’07)
- **Bottom-up search**: Start with bottom clause, remove predicates (ProGolem, Muggleton’09)

Example (Bottom Clause (A))

red(A), big(A), round(A), sibling(A, B), red(B), big(B), round(B)
Bottom-Up Search Advantages

- Omitted Variable Problem
- Not considering a relevant variable
- Bottom-up starts with all attributes

- Myopia Effect
- Top-down search assumes literals conditionally independent given target class
- If features highly correlated, searches very similar hypotheses
Bottom-Up Search Advantages

- **Omitted Variable Problem**
- Not considering a relevant variable
- Bottom-up starts with all attributes

- **Myopia Effect**
- Top-down search assumes literals conditionally independent given target class
- If features highly correlated, searches very similar hypotheses
Non-Determinacy and Recall

Example

\[\text{legalName}(Joe, X); \text{parent}(Joe, Y); \text{Sibling}(Joe, Z)\]

Definition

**Predicate Non-Determinacy:** The number of possible solutions of a given predicate

**Determinate Predicate:** At most one solution

Definition

**Recall:** Imposed bound on predicate non-determinancy
Non-Determinacy and Recall

Example

\texttt{legalName(Joe, X); parent(Joe, Y); sibling(Joe, Z)}

Definition

**Predicate Non-Determinacy:** The number of possible solutions of a given predicate

**Determinate Predicate:** At most one solution

Definition

**Recall:** Imposed bound on predicate non-determinacy
Altering ProGolem Recall

Hexose data:
- highly correlated
- highly non-determinate

Exponential learning time for bottom-up learner

ProGolem: limit bottom clause to first recall instantiations

Placement Bias

1. Default, protein primary sequence
2. Randomize recall selection
3. Domain-dependent, order by distance to binding center
Altering ProGolem Recall

- Hexose data:
  - highly correlated
  - highly non-determinate
- Exponential learning time for bottom-up learner
- ProGolem: limit bottom clause to first *recall* instantiations
- Placement Bias

1. Default, protein primary sequence
2. Randomize recall selection
3. Domain-dependent, order by distance to binding center
Altering ProGolem Recall

- Hexose data:
  - highly correlated
  - highly non-determinate
- Exponential learning time for bottom-up learner
- ProGolem: limit bottom clause to first *recall* instantiations
- Placement Bias

1. Default, protein primary sequence
2. Randomize recall selection
3. Domain-dependent, order by distance to binding center
## ProGolem Recall Results

<table>
<thead>
<tr>
<th>Accuracy</th>
<th>ProGolem recall selection method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>Default</td>
</tr>
<tr>
<td>Mean</td>
<td>59.4%</td>
</tr>
</tbody>
</table>

- Randomized-ProGolem should be used as default
- Recall setting is domain-dependent

**ProGolem insight:**
- Aromatic sandwich
- Novel dependency over LEU and CYS
# ProGolem Recall Results

<table>
<thead>
<tr>
<th>Accuracy</th>
<th>Default</th>
<th>Randomized</th>
<th>Domain-dependent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>59.4%</td>
<td>68.8%</td>
<td>74.8%</td>
</tr>
</tbody>
</table>

- Randomized-ProGolem should be used as default
- Recall setting is domain-dependent

**ProGolem insight:**
- Aromatic sandwich
- Novel dependency over LEU and CYS


Prediction of Protein-Glucose Binding Sites Using SVMs.  

An ILP Approach to Validate Hexose Binding Biochemical Knowledge.  
Outline

1. Differential Prediction
2. Expert Driven Approach
   - Expert Driven Method
   - ProGolem Recall *
3. Model Filtering Approach
4. Differential Prediction Search Approach
5. BI-RADS Information Extraction *
6. Other Work *
Breast-Cancer Stages

In-Situ Cancer Stage

Basement Membrane

Abnormal cells
Breast-Cancer Stages

Invasive Cancer Stage

- Basement Membrane
- Abnormal cells
Age Matters

- Stratify our data (Nichols’06):
  - Younger: < 50 years, pre-menopausal
  - Middle: [50, 65) years, peri-menopausal
  - Older: ≥ 65 years, post-menopausal
- Apply logistic regression on older and younger
- Uncover a differential ability in predicting invasive and in-situ cancer in older vs. younger women
Differential Rules

- Palpable lump => invasive in older
  Recurrence + BI-RADS increase => in situ in younger
  - Younger: rapid proliferation, poor differentiation
    In Situ more likely to be palpable in younger
  - Older: slow tumor growth
    When big enough to be palpable, almost certainly invasive

- Previous invasive biopsy => invasive in older
  - Longer life-span of older women
  - Higher recurrence of invasive tumors
Methodology

- Use tuning sets
- Differential prediction rule if:
  - Target stratum precision > 60% and recall > 10%
  - Compare tuning sets, significantly worse precision on other stratum
- No significant older-specific in situ differential rule
## Middle-Cohort Precision Comparison

Comparing Middle Cohort with:

<table>
<thead>
<tr>
<th>Rule</th>
<th>Older Cohort (p-value)</th>
<th>Younger Cohort (p-value)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Invasive in Older Rules</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rule 1</td>
<td>0.04*</td>
<td>0.50</td>
</tr>
<tr>
<td>Rule 2</td>
<td>0.01*</td>
<td>0.32</td>
</tr>
<tr>
<td>Rule 3</td>
<td>0.05</td>
<td>0.49</td>
</tr>
<tr>
<td>Rule 4</td>
<td>0.26</td>
<td>0.00*</td>
</tr>
<tr>
<td>Rule 5</td>
<td>0.48</td>
<td>0.00*</td>
</tr>
<tr>
<td><strong>In-Situ in Older Rules</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rule 1</td>
<td>0.27</td>
<td>0.06</td>
</tr>
<tr>
<td><strong>Invasive in Younger Rules</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rule 1</td>
<td>0.00*</td>
<td>0.12</td>
</tr>
<tr>
<td><strong>In-Situ in Younger Rules</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rule 1</td>
<td>0.10</td>
<td>0.06</td>
</tr>
</tbody>
</table>

* Statistically significant at the 95% confidence level.


Uncovering Age-Specific Invasive and DCIS Breast Cancer Rules Using ILP.

Outline

1. Differential Prediction
2. Expert Driven Approach
   - Expert Driven Method
   - ProGolem Recall *
3. Model Filtering Approach
4. Differential Prediction Search Approach
5. BI-RADS Information Extraction *
6. Other Work *
Differential Prediction Search Method

Older Stratum Reports → DP Sensitive ILP Classifier → Stratum-Specific Invasive/In Situ Rules

Younger Stratum Reports → DP Sensitive ILP Classifier
**Definition (DP-Sensitive Scoring Function)**

Let $R$ be a 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 $D_o$.

**Example**

\[
Q(R\|D_t, D_o) = S(R\|D_t) - S(R\|D_o)
\]
Definition (DP-Sensitive Scoring Function)

Let $R$ be a rule over the set of instances $X$, and let $D$ 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 $D_o$.

Example

$$Q(R|D_t, D_o) = S(R|D_t) - S(R|D_o)$$
Baseline Method

- Include stratifying attribute as a predicate $p$
- Run ILP over whole dataset
- Select rules containing the predicate $p$
- Rules specific to the stratum the predicate $p$ refers to

Example:

$P(X) \text{ if } \text{red}(X) \land \text{big}(X)$

Houssam Nassif  Differential Relational Learning
Baseline Method

- Include stratifying attribute as a predicate $p$
- Run ILP over whole dataset
- Select rules containing the predicate $p$
- Rules specific to the stratum the predicate $p$ refers to

Example

$P(X)$ if $\text{red}(X) \land \text{big}(X)$
Baseline Method

- Include stratifying attribute as a predicate $p$
- Run ILP over whole dataset
- Select rules containing the predicate $p$
- Rules specific to the stratum the predicate $p$ refers to

Example

$P(X) \text{ if } red(X) \land big(X)$
Michalski Trains (Larson’77)

TRAINS GOING EAST

1. [Diagram]
2. [Diagram]
3. [Diagram]
4. [Diagram]

(a) East A trains: short infront of long; jagged-roof

TRAINS GOING EAST

1. [Diagram]
2. [Diagram]
3. [Diagram]
4. [Diagram]

(b) East B trains: short infront of long; double-hulled

TRAINS GOING WEST

5. [Diagram]
6. [Diagram]
7. [Diagram]
8. [Diagram]
Michalski Trains Experiment

- Size: 100, 1000 (per class and stratum)
- Data: clean, noisy (5% swap)
- Scenarios: one, up to 5 target rules
- Common rules: 1-5

- Rank theory rules by score
- Match rules to target ground truth rules
- PR curve on recovered rules
Michalski Trains Experiment

- Size: 100, 1000 (per class and stratum)
- Data: clean, noisy (5% swap)
- Scenarios: one, up to 5 target rules
- Common rules: 1-5

- Rank theory rules by score
- Match rules to target ground truth rules
- PR curve on recovered rules
Michalski Trains Results

Mean AUC PR for 30 experiments in each block

<table>
<thead>
<tr>
<th>Size</th>
<th>BASE</th>
<th>MF</th>
<th>DPS</th>
<th>BASE</th>
<th>MF</th>
<th>DPS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>clean</td>
<td></td>
<td></td>
<td>noisy</td>
<td></td>
<td></td>
</tr>
<tr>
<td>One target rule scenario</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>0.73</td>
<td><strong>0.83</strong></td>
<td>0.62</td>
<td>0.57</td>
<td><strong>0.62</strong></td>
<td>0.54</td>
</tr>
<tr>
<td>1000</td>
<td>0.87</td>
<td><strong>0.90</strong></td>
<td>0.88</td>
<td>0.63</td>
<td>0.80</td>
<td><strong>0.87</strong></td>
</tr>
<tr>
<td>Multiple target rules scenario</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>0.61</td>
<td><strong>0.70</strong></td>
<td>0.42</td>
<td>0.38</td>
<td><strong>0.52</strong></td>
<td>0.31</td>
</tr>
<tr>
<td>1000</td>
<td>0.75</td>
<td><strong>0.86</strong></td>
<td>0.77</td>
<td>0.52</td>
<td>0.55</td>
<td><strong>0.65</strong></td>
</tr>
</tbody>
</table>

Differential Prediction Search Method

DPS more appropriate for real-world (large + noisy) data
Mean AUC PR for 30 experiments in each block

<table>
<thead>
<tr>
<th>Size</th>
<th>BASE</th>
<th>MF</th>
<th>DPS</th>
<th>BASE</th>
<th>MF</th>
<th>DPS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>clean</td>
<td></td>
<td></td>
<td>noisy</td>
<td></td>
<td></td>
</tr>
<tr>
<td>One target rule scenario</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>0.73</td>
<td><strong>0.83</strong></td>
<td>0.62</td>
<td>0.57</td>
<td><strong>0.62</strong></td>
<td>0.54</td>
</tr>
<tr>
<td>1000</td>
<td>0.87</td>
<td><strong>0.90</strong></td>
<td>0.88</td>
<td>0.63</td>
<td>0.80</td>
<td><strong>0.87</strong></td>
</tr>
<tr>
<td>Multiple target rules scenario</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>0.61</td>
<td><strong>0.70</strong></td>
<td>0.42</td>
<td>0.38</td>
<td><strong>0.52</strong></td>
<td>0.31</td>
</tr>
<tr>
<td>1000</td>
<td>0.75</td>
<td><strong>0.86</strong></td>
<td>0.77</td>
<td>0.52</td>
<td>0.55</td>
<td><strong>0.65</strong></td>
</tr>
</tbody>
</table>

- DPS more appropriate for real-world (large + noisy) data
Revisit In Situ in Older

- Use statistical test, not tuning set
- DP rule must have significantly better precision and recall

- Baseline: No returned DP rule
- MF: Calcification
  - Tumor indolent in older women
  - Asymptomatic in situ detected due to micro-calcifications
  - Novel finding
Revisit In Situ in Older

- Use statistical test, not tuning set
- DP rule must have significantly better precision and recall
- Baseline: No returned DP rule
- MF: Calcification
  - Tumor indolent in older women
  - Asymptomatic in situ detected due to micro-calcifications
  - Novel finding
DPS Older-specific In Situ Rules

1. Calcification
2. Class 2 breast density
   - Lower breast density increases mammogram sensitivity, easier micro-calcification detection
3. Prior in situ biopsy
   - Tumor indolent in older women
4. BI-RADS score increase
5. Screening visit
   - Regular screening age > 40
**Uplift Curve**

**Lift** : Nb of positives in top ranking fraction $p$

**Uplift** : $p \in [0, 1]$, plot $\{p, \text{Lift}_t - \text{Lift}_o\}$

Use theory to form TAN classifier to assign example probability
Logical Differential Prediction Bayes Net
H. Nassif, V. Santos Costa, E.S. Burnside, and D. Page.
Relational Differential Prediction.

H. Nassif, Y. Wu, D. Page, and E.S. Burnside.
Logical Differential Prediction Bayes Net, Improving Breast Cancer Diagnosis for Older Women.
*AMIA*, Chicago, 2012. Accepted.
Outline

1. Differential Prediction
2. Expert Driven Approach
   - Expert Driven Method
   - ProGolem Recall *
3. Model Filtering Approach
4. Differential Prediction Search Approach
5. BI-RADS Information Extraction *
6. Other Work *
Information from Lexicon

- Lexicon specifies synonyms
  - E.g.: Equal density, Isodense
- Lexicon allows for ambiguous wording

<table>
<thead>
<tr>
<th>Text</th>
<th>Concept</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>indistinct</strong> margin</td>
<td>indistinct margin</td>
</tr>
<tr>
<td><strong>indistinct</strong> calcification</td>
<td>amorphous calcification</td>
</tr>
<tr>
<td><strong>indistinct</strong> image</td>
<td>not a BI-RADS concept</td>
</tr>
</tbody>
</table>
Algorithm Flowchart

1. BI-RADS Reports → Syntax Analyzer
2. Experts
3. Lexicon
4. Concept Finder (Semantic Parser)
5. Negation Detector (Lexical Scanner)
6. BI-RADS Features

- Regular expression
- Straight-forward negation
- Negation-deactivation triggers
Rule Generation Example

- Aim: Skin Thickening concept
- Lexicon specifies “skin thickening”
- Try “skin” and “thickening” in same sentence
  - thickening of the overlying skin
  - marker placed on the skin overlying a palpable focal area of thickening in the upper outer right breast
- Experts suggest “skin” and “thickening” in close proximity

- Start with a large scope
  - Assess number of true and false positives
- Move to smaller scopes
  - Assess number of false negatives
- Experts decide on the best distance
Rule Generation Example

- **Aim:** Skin Thickening concept
- **Lexicon specifies** “skin thickening”
- **Try** “skin” and “thickening” in same sentence
  - **thickening** of the overlying **skin**
  - marker placed on the **skin** overlying a palpable focal area of **thickening** in the upper outer right breast
- **Experts suggest** “skin” and “thickening” in close proximity

- **Start with a large scope**
  - Assess number of true and false positives
- **Move to smaller scopes**
  - Assess number of false negatives
- **Experts decide** on the best distance
BI-RADS Information Extraction Results

- Outperforms manual extraction
- Cross-institution portability
- Extends to other languages

- First successful BI-RADS features extractor
- First breast tissue composition extractor
- First Portuguese BI-RADS features extractor
B. Percha, **H. Nassif**, J. Lipson, E. Burnside, and D. Rubin.
Automatic Classification of Mammography Reports by BI-RADS Breast Tissue Composition Class.
*JAMIA*, Online First, 2012.

Extracting BI-RADS Features from Portuguese Clinical Texts.

Information Extraction for Clinical Data Mining: A Mammography Case Study.
Outline

1. Differential Prediction
2. Expert Driven Approach
   - Expert Driven Method
   - ProGoIem Recall *
3. Model Filtering Approach
4. Differential Prediction Search Approach
5. BI-RADS Information Extraction *
6. Other Work *
Other Work

- SAYU
- Mammography upgrades

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

- Differential Prediction for Adverse Drug Events
Summary

- Relational differential prediction
- Introduce differential predictive rules
  - Expert Driven approach
  - Model Filtering approach
  - Differential Predictive Search approach
- Recommend DPS for real world applications
- Glucose classifier, hexose data-driven validation
- Randomize ProGolem recall
- Logical Differential Prediction Bayes Net
- English and Portuguese BI-RADS features extraction
Acknowledgments

- Adviser David Page
- Grant PI Elizabeth Burnside
- Committee members
- Research collaborators and lab mates
- Talk attendance
- Family and friends
- Mother-in-law Nawal
- My wonderful Carole
Acknowledgments

- Adviser David Page
- Grant PI Elizabeth Burnside
- Committee members
- Research collaborators and lab mates
- Talk attendance
- Family and friends
- Mother-in-law Nawal
- My wonderful Carole
Outline

7 Appendix A: Hexoses
- Atomic Interactions
- Hexose Features

8 Appendix B: Algorithms
- RF-SVM
- ILP
- Instance Relabeling
Covalent Bonds

- Close and strong interaction
- Forms a molecule
- Atoms share electrons
- Electronegativity:
  - Equal $\Rightarrow$ nonpolar
  - Different $\Rightarrow$ polar
- Partial charges

Electronegativity: Measure of atom’s attraction for electrons
Covalent Bonds

- Close and strong interaction
- Forms a molecule
- Atoms share electrons
- Electronegativity:
  - Equal $\Rightarrow$ nonpolar
  - Different $\Rightarrow$ polar
- Partial charges

**Definition**

**Electronegativity:** Measure of atom’s attraction for electrons
Covalent Bonds

- Close and strong interaction
- Forms a molecule
- Atoms share electrons
- Electronegativity:
  - Equal $\Rightarrow$ nonpolar
  - Different $\Rightarrow$ polar
- Partial charges

Definition

**Electronegativity:** Measure of atom’s attraction for electrons
Hydrogen Bonds

- Attraction between a positively charged H and a negatively charged atom
- Hexose attaches to the protein using hydrogen bonds

Hydrogen bond

\[
\begin{align*}
\text{H} & \quad \text{O} \quad \text{H}^\delta \quad \text{N}^{-\delta} \quad \text{H} \\
\text{H} & \quad \text{H} \\
\end{align*}
\]
Van der Waals and Hydrophobicity

**Definition**

**Van der Waals Forces:** Weak electrostatic attraction and repulsion forces

**Definition (Hydrophobicity)**


- Dual nature:
  - Pyranose ring is hydrophobic
  - Hydroxyl group is hydrophilic
Van der Waals and Hydrophobicity

**Definition**

*Van der Waals Forces*: Weak electrostatic attraction and repulsion forces

**Definition (Hydrophobicity)**

*Hydrophobic*: water hating. *Hydrophilic*: water loving.
Hydrophobic/Hydrophilic atoms tend to gather together.

- Dual nature:
  - Pyranose ring is hydrophobic
  - Hydroxyl group is hydrophilic
Appendix A: Hexoses

- Atomic Interactions
- Hexose Features

Appendix B: Algorithms

- RF-SVM
- ILP
- Instance Relabeling
# Hexose Features

<table>
<thead>
<tr>
<th>Atomic Feature</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Charge</td>
<td>Negative, Neutral, Positive</td>
</tr>
<tr>
<td>Hydrogen-bonding</td>
<td>Non-hydrogen bonding, Hydrogen-bonding</td>
</tr>
<tr>
<td>Hydrophobicity</td>
<td>Hydrophilic, Hydroneutral, Hydrophobic</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Residue Grouping</th>
<th>Amino Acids</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aromatic</td>
<td>HIS, PHE, TRP, TYR</td>
</tr>
<tr>
<td>Aliphatic</td>
<td>Ala, Ile, Leu, Met, Val</td>
</tr>
<tr>
<td>Neutral</td>
<td>ASN, CYS, GLN, GLY, PRO, SER, THR</td>
</tr>
<tr>
<td>Acidic</td>
<td>ASP, GLU</td>
</tr>
<tr>
<td>Basic</td>
<td>ARG, LYS</td>
</tr>
</tbody>
</table>
## Atomic Chemical Properties I

<table>
<thead>
<tr>
<th>PDB atom symbol</th>
<th>Residues</th>
<th>Partial Charge</th>
<th>Hydrophobicity</th>
<th>Hydrogen Bonding</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Amino acid oxygen atoms</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>O</td>
<td>All amino acids</td>
<td>0</td>
<td>HPHIL</td>
<td>HB</td>
</tr>
<tr>
<td>OXT</td>
<td>All amino acids</td>
<td>-ve</td>
<td>HPHIL</td>
<td>HB</td>
</tr>
<tr>
<td>OE1, OE2, OD1, OD2</td>
<td>GLU, ASP</td>
<td>-ve</td>
<td>HPHIL</td>
<td>HB</td>
</tr>
<tr>
<td>OE1, OD1</td>
<td>GLN, ASN</td>
<td>0</td>
<td>HPHIL</td>
<td>HB</td>
</tr>
<tr>
<td>OG, OG1, OH</td>
<td>SER, THR, TYR</td>
<td>0</td>
<td>HPHIL</td>
<td>HB</td>
</tr>
<tr>
<td><strong>Amino acid carbon atoms</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>All amino acids</td>
<td>0</td>
<td>HNEUT</td>
<td>NHB</td>
</tr>
<tr>
<td>CA</td>
<td>All amino acids</td>
<td>0</td>
<td>HNEUT</td>
<td>NHB</td>
</tr>
<tr>
<td>CB, CG, CD, CE</td>
<td>ALA, SER, THR, CYS, ASP, ASN, GLU, GLN, ARG, LYS, PRO</td>
<td>0</td>
<td>HNEUT</td>
<td>NHB</td>
</tr>
<tr>
<td>CB, CG, CD, CE</td>
<td>LEU, VAL, ILE, MET</td>
<td>0</td>
<td>HPHOB</td>
<td>NHB</td>
</tr>
<tr>
<td>CG1, CG2, CD1, CD2, CD1</td>
<td>LEU, VAL, ILE</td>
<td>0</td>
<td>HPHOB</td>
<td>NHB</td>
</tr>
<tr>
<td>CG, CD1, CD2, CE1, CE2, CZ, CG,CD1, CD2, CE2, CE3, CZ2, CZ3, CH2</td>
<td>PHE, TYR, TRP</td>
<td>0</td>
<td>HPHOB</td>
<td>NHB</td>
</tr>
<tr>
<td>CG, CD2, CE1</td>
<td>HIS</td>
<td>0</td>
<td>HPHOB</td>
<td>NHB</td>
</tr>
</tbody>
</table>
## Atomic Chemical Properties II

<table>
<thead>
<tr>
<th>PDB atom symbol</th>
<th>Residues</th>
<th>Partial Charge</th>
<th>Hydrophobicity</th>
<th>Hydrogen Bonding</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Amino acid nitrogen atoms</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>All amino acids except PRO</td>
<td>0</td>
<td>HPHIL</td>
<td>HB</td>
</tr>
<tr>
<td>N</td>
<td>PRO</td>
<td>0</td>
<td>HPHIL</td>
<td>NHB</td>
</tr>
<tr>
<td>NE2, ND2</td>
<td>GLN, ASN</td>
<td>0</td>
<td>HPHIL</td>
<td>HB</td>
</tr>
<tr>
<td>NZ</td>
<td>LYS</td>
<td>+ve</td>
<td>HPHIL</td>
<td>HB</td>
</tr>
<tr>
<td>NE</td>
<td>ARG</td>
<td>+ve</td>
<td>HPHIL</td>
<td>NHB</td>
</tr>
<tr>
<td>NH1, NH2</td>
<td>ARG</td>
<td>+ve</td>
<td>HPHIL</td>
<td>HB</td>
</tr>
<tr>
<td>ND1, NE2</td>
<td>HIS</td>
<td>0</td>
<td>HPHIL</td>
<td>HB</td>
</tr>
<tr>
<td>NE1</td>
<td>TRP</td>
<td>0</td>
<td>HNEUT</td>
<td>NHB</td>
</tr>
<tr>
<td><strong>Amino acid sulfur atoms</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SG</td>
<td>CYS</td>
<td>0</td>
<td>HPHIL</td>
<td>HB</td>
</tr>
<tr>
<td>SD</td>
<td>MET</td>
<td>0</td>
<td>HNEUT</td>
<td>NHB</td>
</tr>
<tr>
<td><strong>Water and ions atoms</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>O</td>
<td>HOH</td>
<td>0</td>
<td>HPHIL</td>
<td>HB</td>
</tr>
<tr>
<td>O1, O2, O3, O4</td>
<td>SO4, 2HP</td>
<td>-ve</td>
<td>HPHIL</td>
<td>HB</td>
</tr>
<tr>
<td>CA, MG, ZN</td>
<td>CA, MG, ZN</td>
<td>+ve</td>
<td>HPHIL</td>
<td>HB</td>
</tr>
</tbody>
</table>
Outline

Appendix A: Hexoses
- Atomic Interactions
- Hexose Features

Appendix B: Algorithms
- RF-SVM
- ILP
- Instance Relabeling
Support Vector Machines (SVM, Vapnik’98)

- Construct the *optimal separating hyperplane* (usually in a higher feature space)
  - Maximize *margins*: minimal distance from the hyperplane
  - Only *Support Vectors (SV)* specify the margins/hyperplane
  - Small number of SV $\iff$ good generalization
Support Vector Machines (SVM, Vapnik’98)

- Construct the *optimal separating hyperplane* (usually in a higher feature space)
- Maximize *margins*: minimal distance from the hyperplane
- Only *Support Vectors (SV)* specify the margins/hyperplane
- Small number of SV ⇔ good generalization
Support Vector Machines (SVM, *Vapnik’98*)

- Construct the *optimal separating hyperplane* (usually in a higher feature space)
- Maximize *margins*: minimal distance from the hyperplane
- Only *Support Vectors (SV)* specify the margins/hyperplane
- Small number of SV $\Leftrightarrow$ good generalization
Support Vector Machines (SVM, Vapnik’98)

- Construct the *optimal separating hyperplane* (usually in a higher feature space)
- Maximize *margins*: minimal distance from the hyperplane
- Only *Support Vectors (SV)* specify the margins/hyperplane
- Small number of SV $\Leftrightarrow$ good generalization
Random Forest (RF, Breiman’01)

- High features/examples ratio $\Rightarrow$ curse of dimensionality
- Feature selection: select the best feature subset

Random Forest feature selection:
- Based on multiple classification trees
- Provides direct feature importance measure
- Can be used when feature number $\gg$ samples
- Robust to noise
- Low bias and low variance
Random Forest (RF, *Breiman’01*)

- High features/examples ratio $\Rightarrow$ *curse of dimensionality*
- *Feature selection*: select the best feature subset

Random Forest feature selection:
- Based on multiple classification trees
- Provides direct feature importance measure
- Can be used when feature number $\gg$ samples
- Robust to noise
- Low bias and low variance
RF Feature Importance Score (Díaz-Uriarte’06)

- Create $j$ bootstrap datasets (select $n$ with replacement)
- Out-of-bag (OOB): $\approx 1/3$ of items not included
- Grow a decision tree over each dataset
  - At each tree node, select $q$ features randomly
  - Split node according to best split among the $q$ features
  - Each tree remains unpruned (low-bias)
- Let the tree classify its own OOB data
- Compute the number of correctly classified samples
- Permute the values of feature $k$ in the OOB
- Classify modified OOB, compute classification difference
- Feature Importance Score: Resulting accuracy decrease
Appendix A: Hexoses

- Atomic Interactions
- Hexose Features

Appendix B: Algorithms

- RF-SVM
- ILP
- Instance Relabeling
Aleph (Top-Down, Srinivasan’07)

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

1: 
2: $Learned\_rules \leftarrow \{\}$
3: $Pos \leftarrow$ all positive examples in $E$
4: **while** $Pos$ **do**
5: Select example $e \in Pos$
6: Construct bottom clause $\bot_e$ from $e$, $M$ and $B$ ▶ Saturation step
7: $Candidate\_literals \leftarrow \text{Literals}(\bot_e)$
8: $New\_rule \leftarrow \text{pos}(X)$ ▶ Most general rule
9: **repeat** ▶ Top-down reduction step
10: $Best\_literal \leftarrow \underset{L \in Candidate\_literals}{\text{argmax}} S(\text{New\_rule with precondition } L)$
11: Add $Best\_literal$ to preconditions of $New\_rule$
12: **until** No more $S(\text{New\_rule})$ score improvement
13: $Learned\_rules \leftarrow Learned\_rules + New\_rule$
14: $Pos \leftarrow Pos - \{\text{members of } Pos \text{ covered by } New\_rule\}$
15: **end while**
16: **return** $Learned\_rules$
ProGolem (Bottom-Up, Muggleton’09)

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

1: 
2: $Learned\_rules \leftarrow \{\}$
3: $Pos \leftarrow$ all positive examples in $E$
4: **while** $Pos$ **do**
5: Select example $e \in Pos$
6: Construct bottom clause $\bot_e$ from $e$, $M$ and $B$ $\triangleright$ Saturation step
7: $New\_rule \leftarrow \bot_e$ $\triangleright$ Most specific rule
8: **repeat** $\triangleright$ Bottom-up reduction step
9: Select a different example $e' \in Pos$
10: $Blocking\_literals \leftarrow ARMG(New\_rule, e')$
11: Remove $Blocking\_literals$ from preconditions of $New\_rule$
12: **until** No more $S(New\_rule)$ score improvement
13: $Learned\_rules \leftarrow Learnend\_rules + New\_rule$
14: $Pos \leftarrow Pos - \{\text{members of } Pos \text{ covered by } New\_rule\}$
15: **end while**
16: **return** $Learned\_rules$
Outline

7 Appendix A: Hexoses
   • Atomic Interactions
   • Hexose Features

8 Appendix B: Algorithms
   • RF-SVM
   • ILP
   • Instance Relabeling
Coverage Scoring Function

- Rule coverage score: $\text{Cover}(P) - \text{Cover}(N)$
- DP: $(\text{Cover}(P_1) - \text{Cover}(N_1)) - (\text{Cover}(P_2) - \text{Cover}(N_2))$
Coverage Scoring Function

- Rule coverage score: $\text{Cover}(P) - \text{Cover}(N)$
- DP: $(\text{Cover}(P_1) - \text{Cover}(N_1)) - (\text{Cover}(P_2) - \text{Cover}(N_2))$
Instance Relabeling DP Method \textit{(Page’12)}

- Relabel $Pos = P1 + N2$
- Relabel $Neg = P2 + N1$
- Run standard ILP
- $Cover(Pos) - Cover(Neg)$
- $Cover(P1+N2) - Cover(P2+N1)$
- $(Cover(P1) + Cover(N2)) - (Cover(P2) + Cover(N1))$
- $(Cover(P1) - Cover(N1)) - (Cover(P2) - Cover(N2))$
Relabel $\text{Pos} = P_1 + N_2$
Relabel $\text{Neg} = P_2 + N_1$
Run standard ILP
$\text{Cover}(\text{Pos}) - \text{Cover}(\text{Neg})$
$\text{Cover}(P_1 + N_2) - \text{Cover}(P_2 + N_1)$
$(\text{Cover}(P_1) + \text{Cover}(N_2)) - (\text{Cover}(P_2) + \text{Cover}(N_1))$
$(\text{Cover}(P_1) - \text{Cover}(N_1)) - (\text{Cover}(P_2) - \text{Cover}(N_2))$
Instance Relabeling DP Method (Page’12)

- Relabel $\text{Pos} = P1 + N2$
- Relabel $\text{Neg} = P2 + N1$
- Run standard ILP
- $\text{Cover}(\text{Pos}) - \text{Cover}(\text{Neg})$
- $\text{Cover}(P1+N2) - \text{Cover}(P2+N1)$
- $(\text{Cover}(P1) + \text{Cover}(N2)) - (\text{Cover}(P2) + \text{Cover}(N1))$
- $(\text{Cover}(P1) - \text{Cover}(N1)) - (\text{Cover}(P2) - \text{Cover}(N2))$