Differential Relational Learning

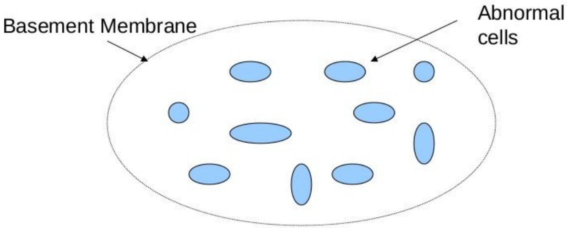
Houssam Nassif

Thesis Defense
03 August 2012

Outline

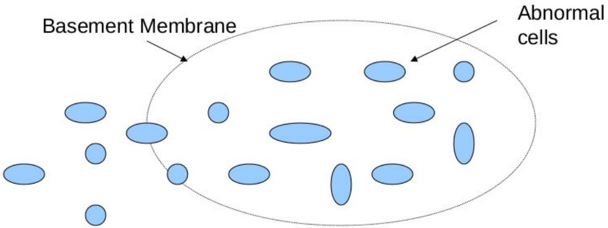
Outline

Breast-Cancer Stages



In-Situ Cancer Stage

Breast-Cancer Stages



Invasive Cancer Stage

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?

Cancer Stage Features

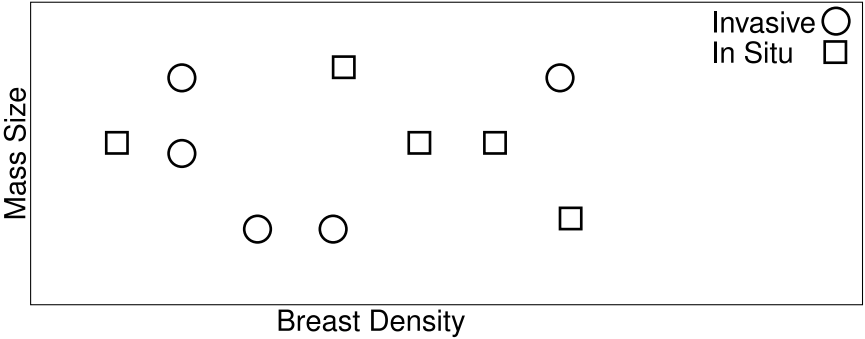
- In Situ can develop into Invasive
  - Current practice: Always treat In Situ
- Time to spread may be very long
  - Patient may die of other causes
  - Over-diagnosis (and over-treatment)

- What features characterize In Situ in older patients?
- What features change between older and younger?

Differential Prediction

Definition (*Cleary'68*)

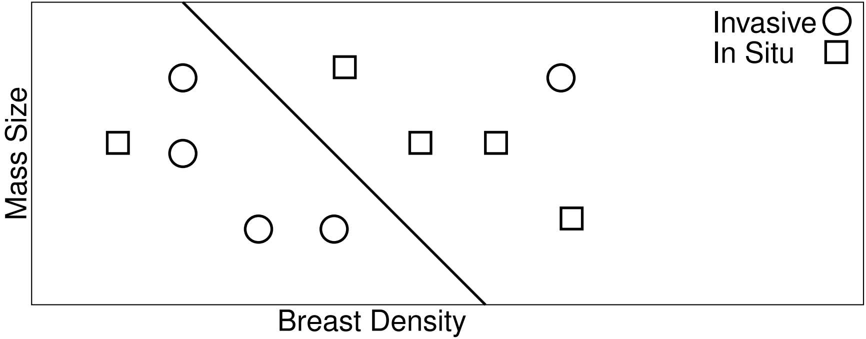
Differential Prediction (DP): case where consistent nonzero errors of prediction are made for members of a given subgroup



Differential Prediction

Definition (*Cleary'68*)

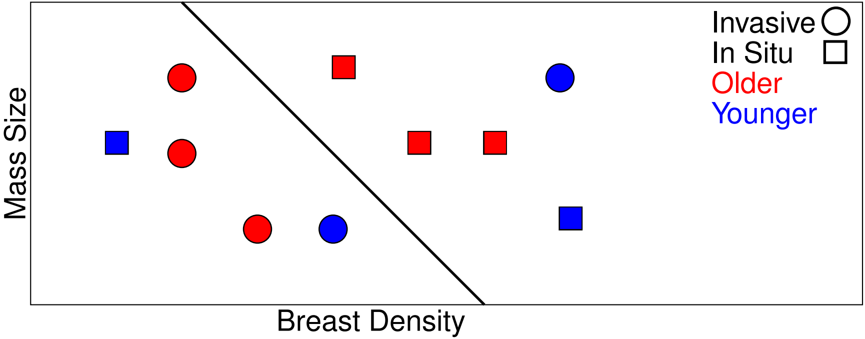
Differential Prediction (DP): case where consistent nonzero errors of prediction are made for members of a given subgroup



Differential Prediction

Definition (*Cleary'68*)

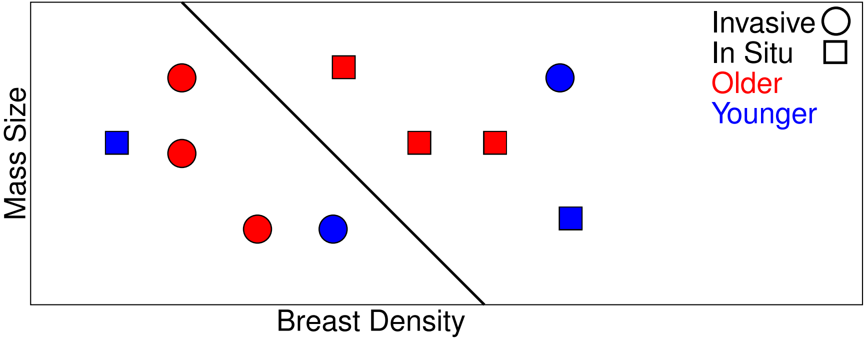
Differential Prediction (DP): case where consistent nonzero errors of prediction are made for members of a given subgroup



Differential Prediction

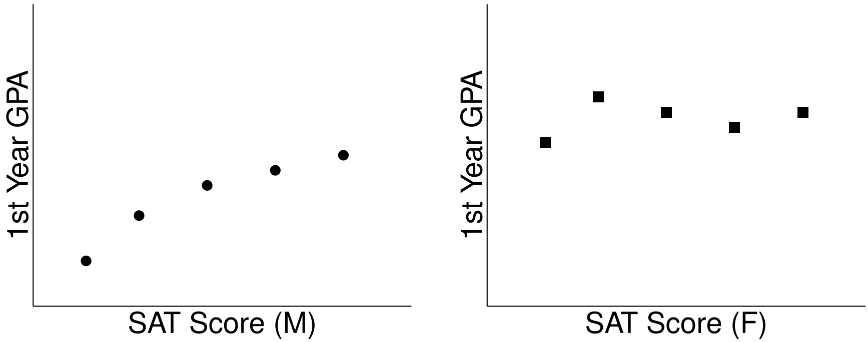
Definition (*Cleary'68*)

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



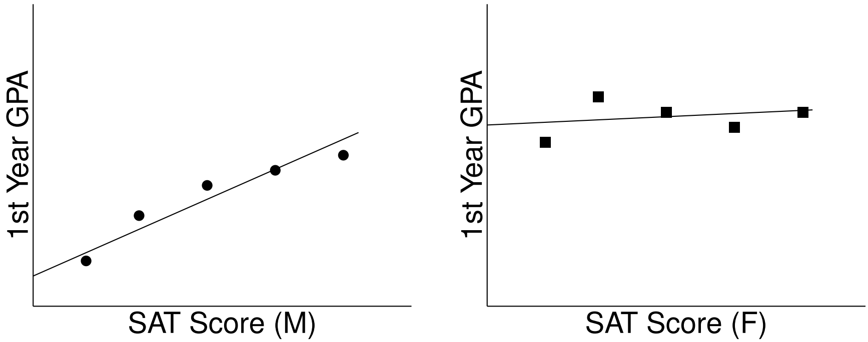
Using Regression to Detect DP

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



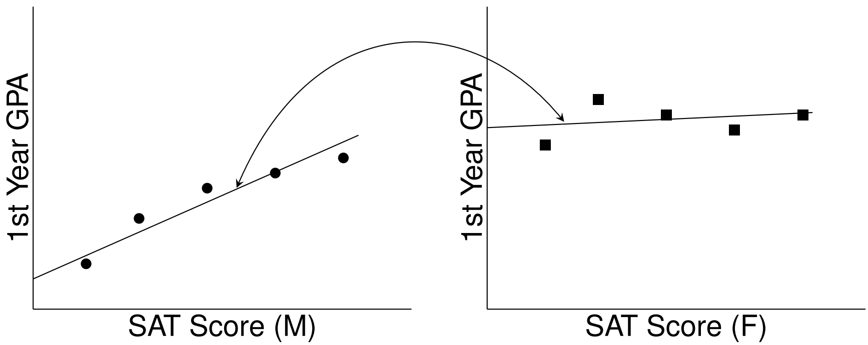
Using Regression to Detect DP

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



Using Regression to Detect DP

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



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

Thesis Statement

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

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:

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

Differential Predictive Rule

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 $\mathcal{D}$ 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:

$$\forall i \neq j,\ S(c_j|D_j) \gg S(c_j|D_i). \quad (2)$$

- 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

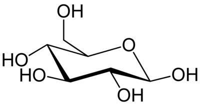
- Expert can be a human or a machine
- Method can be applied to non-rule-learning classifiers

Expert Driven Approach

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

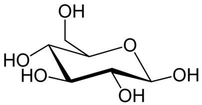
- Expert can be a human or a machine
- Method can be applied to non-rule-learning classifiers

Hexose-Binding Modeling



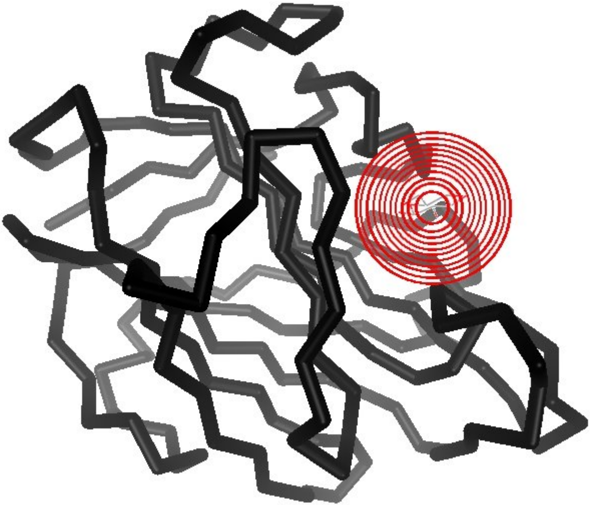
- Galactose, glucose, mannose
- High specificity to diverse protein families
- Interesting to uncover differential binding patterns between glucose-binding and general hexose-binding

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

```
procedure EXTRACTFEATURES(binding site center)
    for all concentric layers do
        for all PDB atoms do
            get distance from center
            get charge
            get hydrophobicity
            get hydrogen-bonding
            get residue
        end for
    end for
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

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

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

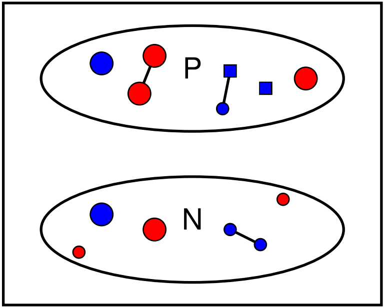
Expert Driven Differential Rules

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

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

Outline

ILP Example

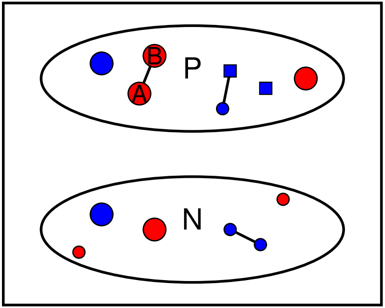


Example

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

- Pick a positive instance
- $P(X)$ *if* $square(X)$
- $P(X)$ *if* $red(X) \wedge big(X)$
  - 1 false positive
- $P(X)$ *if* $sibling(X, Y) \wedge square(Y)$
- 1 false negative
- Form theory

ILP Example

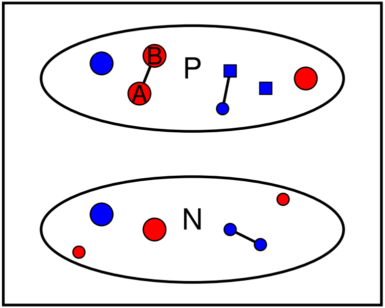


Example

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

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

ILP Example

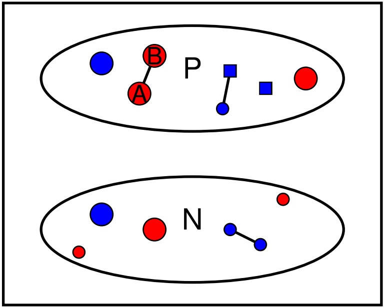


Example

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

- Pick a positive instance
- $P(X)$ *if* $square(X)$
- $P(X)$ *if* $red(X) \wedge big(X)$
  - 1 false positive
- $P(X)$ *if* $sibling(X, Y) \wedge square(Y)$
- 1 false negative
- Form theory

ILP Example

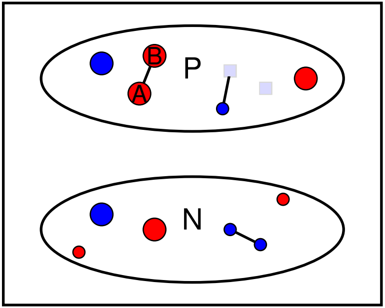


Example

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

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

ILP Example

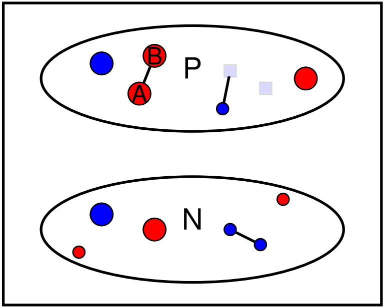


Example

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

- Pick a positive instance
- $P(X)$ *if* $square(X)$
- $P(X)$ *if* $red(X) \wedge big(X)$
  - 1 false positive
- $P(X)$ *if* $sibling(X, Y) \wedge square(Y)$
- 1 false negative
- Form theory

ILP Example

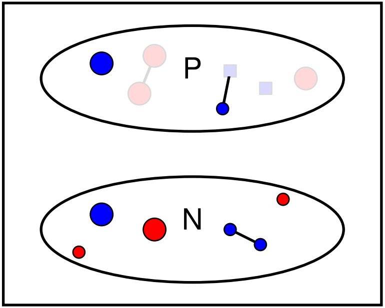


- Pick a positive instance
- $P(X)$ *if* $square(X)$
- $P(X)$ *if* $red(X) \wedge big(X)$
  - 1 false positive
- $P(X)$ *if* $sibling(X, Y) \wedge square(Y)$
- 1 false negative
- Form theory

Example

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

ILP Example

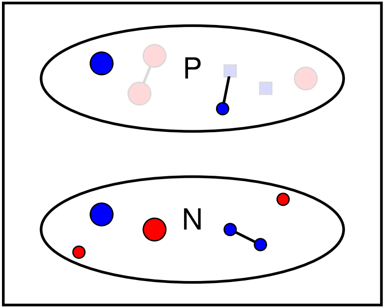


Example

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

- Pick a positive instance
- $P(X)$ *if* $square(X)$
- $P(X)$ *if* $red(X) \wedge big(X)$
  - 1 false positive
- $P(X)$ *if* $sibling(X, Y) \wedge square(Y)$
- 1 false negative
- Form theory

ILP Example

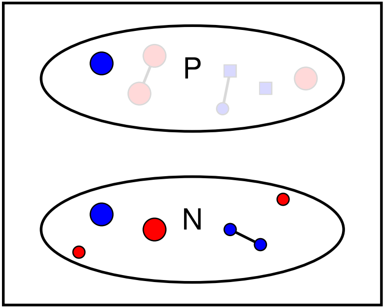


- Pick a positive instance
- $P(X)$ *if* $square(X)$
- $P(X)$ *if* $red(X) \wedge big(X)$
  - 1 false positive
- $P(X)$ *if* $sibling(X, Y) \wedge square(Y)$
- 1 false negative
- Form theory

Example

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

ILP Example

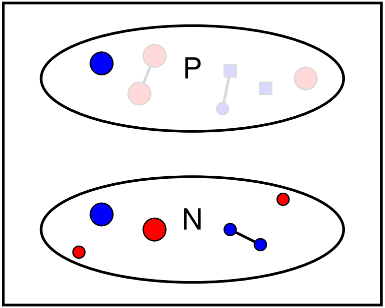


- Pick a positive instance
- $P(X)$ *if* $square(X)$
- $P(X)$ *if* $red(X) \wedge big(X)$
  - 1 false positive
- $P(X)$ *if* $sibling(X, Y) \wedge square(Y)$
- 1 false negative
- Form theory

Example

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

ILP Example

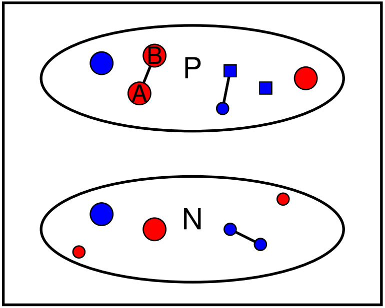


Example

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

- Pick a positive instance
- $P(X)$ *if* $square(X)$
- $P(X)$ *if* $red(X) \wedge big(X)$
  - 1 false positive
- $P(X)$ *if* $sibling(X, Y) \wedge square(Y)$
- 1 false negative
- Form theory

ILP Search

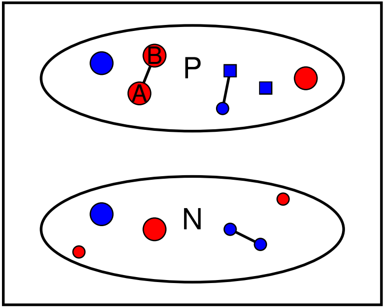


Example (Bottom Clause (A))

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

- 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

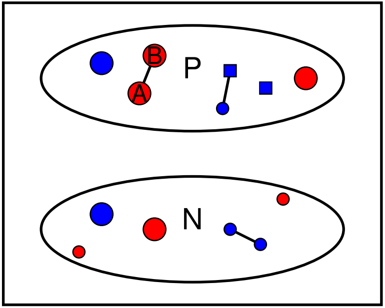


- Pick a positive instance
- Construct the *Bottom Clause*, most specific clause
- *Top-down search*: Start with most general rule, add bottom clause predicates (Aleph, Srinivasan'07)
- *Bottom-up search*: Start with bottom clause, remove predicates (ProGolem, Muggleton'09)

Example (Bottom Clause (A))

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

ILP Search

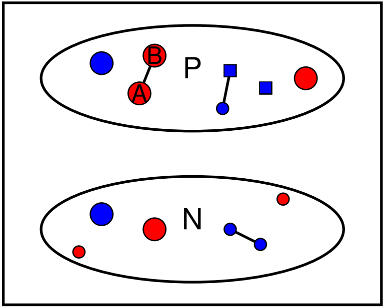


- Pick a positive instance
- Construct the *Bottom Clause*, most specific clause
- *Top-down search*: Start with most general rule, add bottom clause predicates (Aleph, Srinivasan'07)
- *Bottom-up search*: Start with bottom clause, remove predicates (ProGolem, Muggleton'09)

Example (Bottom Clause (A))

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

ILP Search



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)

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

$legalName(Joe, X)$; $parent(Joe, Y)$; $sibling(Joe, Z)$

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

Non-Determinacy and Recall

Example

$legalName(Joe, X); parent(Joe, Y); sibling(Joe, Z)$

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

	ProGolem recall selection method		
Accuracy	Default	Randomized	Domain-dependent
Mean	59.4%	68.8%	74.8%

- 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

	ProGolem recall selection method		
Accuracy	Default	Randomized	Domain-dependent
Mean	59.4%	68.8%	74.8%

- Randomized-ProGolem should be used as default
- Recall setting is domain-dependent
- ProGolem insight:
  - Aromatic sandwich
  - Novel dependency over LEU and CYS

ED Bibliography

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.

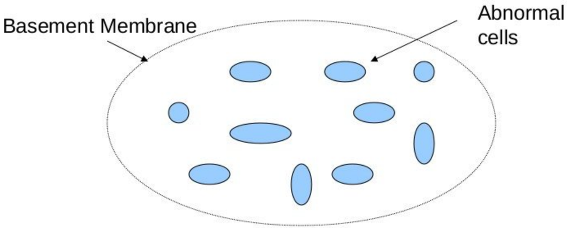
H. Nassif, H. Al-Ali, S. Khuri, and W. Keyrouz.
Prediction of Protein-Glucose Binding Sites Using SVMs.
Proteins, 77(1):121-132, 2009.

H. Nassif, H. Al-Ali, S. Khuri, W. Keyrouz and D. Page.
An ILP Approach to Validate Hexose Binding Biochemical Knowledge.
ILP, Leuven, Belgium, pp. 149-165, 2009.

Outline

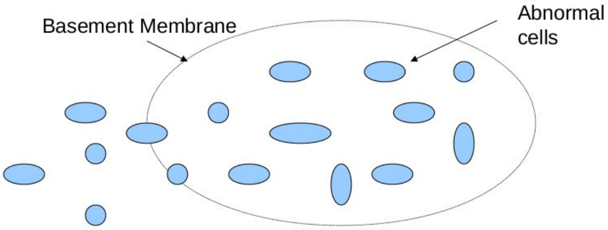
1. Differential Prediction
2. Expert Driven Approach
   - Expert Driven Method
   - ProGolem Recall *
3. Model Filtering Approach
4. Differential Prediction Search Approach
5. BI-RADS Information Extraction *
6. Other Work *

Breast-Cancer Stages



In-Situ Cancer Stage

Breast-Cancer Stages



Invasive Cancer Stage

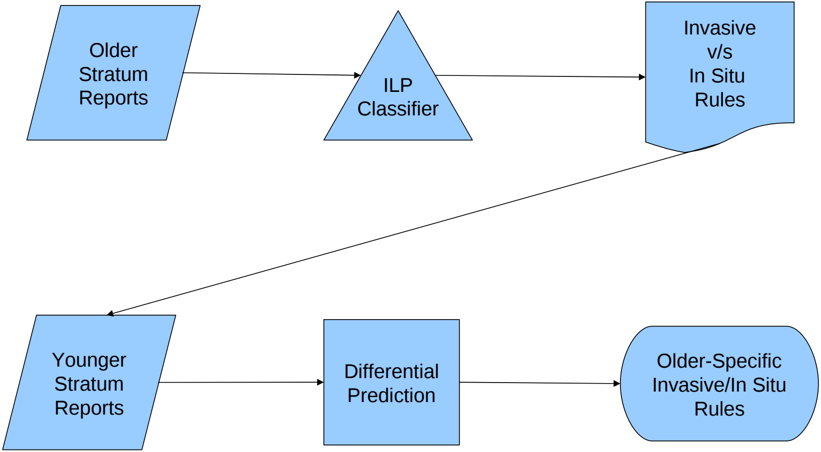
Age Matters

- 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

Model Filtering Method



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:	
Rule	Older Cohort (*p*-value)	Younger Cohort (*p*-value)
Invasive in Older Rules		
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 in Older Rules		
Rule 1	0.27	0.06
Invasive in Younger Rules		
Rule 1	0.00*	0.12
In-Situ in Younger Rules		
Rule 1	0.10	0.06

* Statistically significant at the 95% confidence level.

MF Bibliography

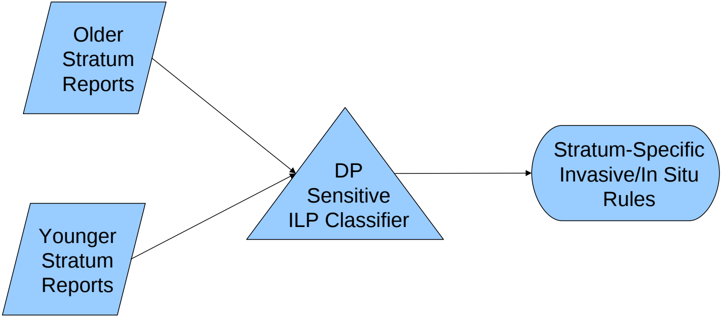
M. Ayvaci, O. Alagoz, J. Chhatwal, A. del Rio, E. Sickles, **H. Nassif**, K. Kerlikowske, and E. Burnside.
Predicting invasive breast cancer versus DCIS in different age groups.
PLoS ONE. Submitted.

H. Nassif, D. Page, M. Ayvaci, J. Shavlik, and E.S. Burnside.
Uncovering Age-Specific Invasive and DCIS Breast Cancer Rules Using ILP.
IHI, Arlington, VA, pp. 76-82, 2010.

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



DP-Sensitive Scoring Function

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

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

DP-Sensitive Scoring Function

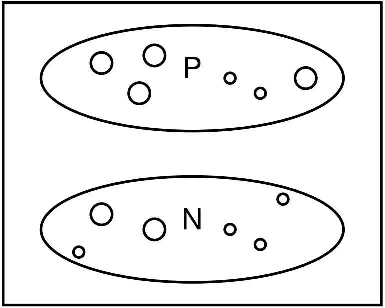
Definition (**DP-Sensitive Scoring Function**)

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

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

Baseline Method

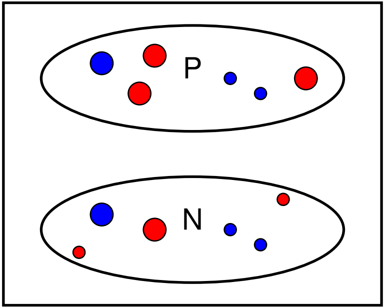


- Include stratifying attribute as a predicate p
- Run ILP over whole dataset
- Select rules containing the predicate p
- Rules specific to the stratum the predicate p refers to

Example

$P(X)$ *if* $red(X) \wedge big(X)$

Baseline Method

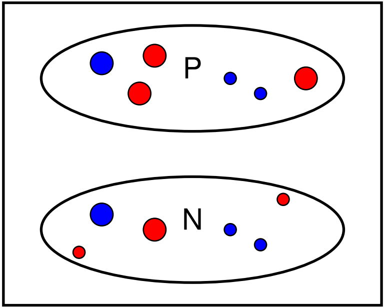


- Include stratifying attribute as a predicate *p*
- Run ILP over whole dataset
- Select rules containing the predicate *p*
- Rules specific to the stratum the predicate *p* refers to

Example

$P(X)$ *if* $red(X) \wedge big(X)$

Baseline Method

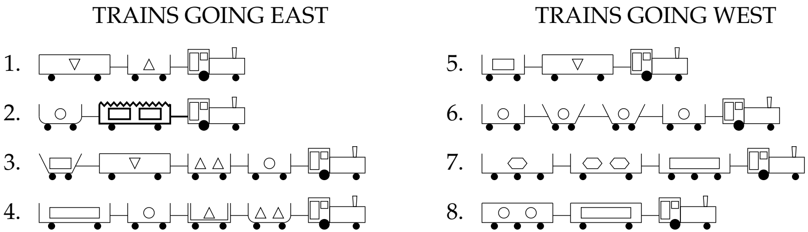


- Include stratifying attribute as a predicate *p*
- Run ILP over whole dataset
- Select rules containing the predicate *p*
- Rules specific to the stratum the predicate *p* refers to

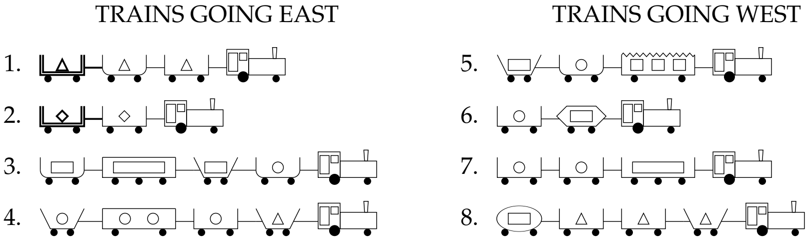
Example

$P(X)$ *if* $red(X) \wedge big(X)$

Michalski Trains (*Larson'77*)



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



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

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

Size	clean BASE	clean MF	clean DPS	noisy BASE	noisy MF	noisy DPS
One target rule scenario						
100	0.73	**0.83**	0.62	0.57	**0.62**	0.54
1000	0.87	**0.90**	0.88	0.63	0.80	**0.87**
Multiple target rules scenario						
100	0.61	**0.70**	0.42	0.38	**0.52**	0.31
1000	0.75	**0.86**	0.77	0.52	0.55	**0.65**

- DPS more appropriate for real-world (large + noisy) data

Michalski Trains Results

Mean AUC PR for 30 experiments in each block

Size	clean BASE	clean MF	clean DPS	noisy BASE	noisy MF	noisy DPS
One target rule scenario						
100	0.73	**0.83**	0.62	0.57	**0.62**	0.54
1000	0.87	**0.90**	0.88	0.63	0.80	**0.87**
Multiple target rules scenario						
100	0.61	**0.70**	0.42	0.38	**0.52**	0.31
1000	0.75	**0.86**	0.77	0.52	0.55	**0.65**

- 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

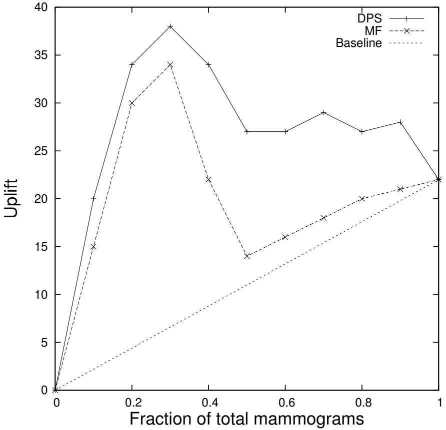
1. Calcification
2. Class 2 breast density
   - Lower breast density increases mammogram sensitivity, easier micro-calcification detection
3. Prior in situ biopsy
   - Tumor indolent in older women
4. BI-RADS score increase
5. Screening visit
   - Regular screening age > 40

Uplift Curve

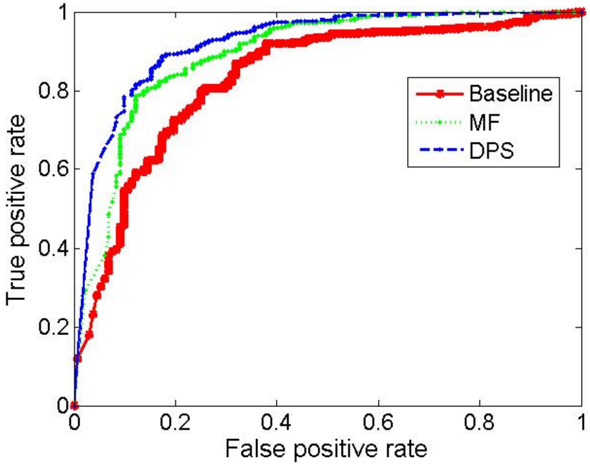
Lift : Nb of positives in top ranking fraction p

Uplift : $p \in [0,1]$, plot $\{p, Lift_t - Lift_o\}$

Use theory to form TAN classifier to assign example probability



Logical Differential Prediction Bayes Net

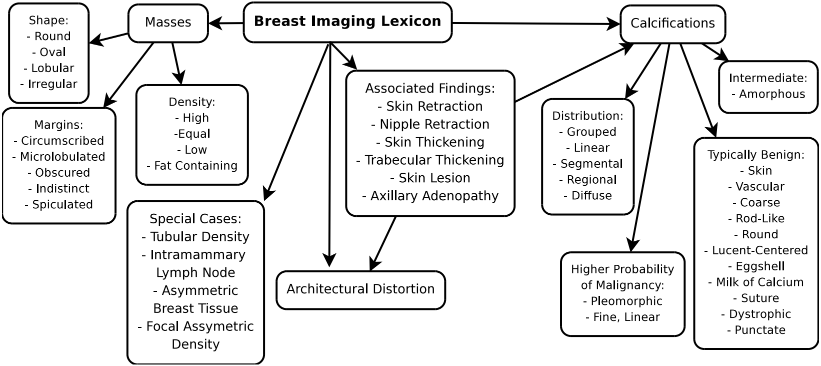


DPS Bibliography

- **H. Nassif**, V. Santos Costa, E.S. Burnside, and D. Page.
  Relational Differential Prediction.
  ECML, Bristol, UK, pp. 617-632, 2012.
- **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

Breast Imaging Reporting & Data System (BI-RADS)

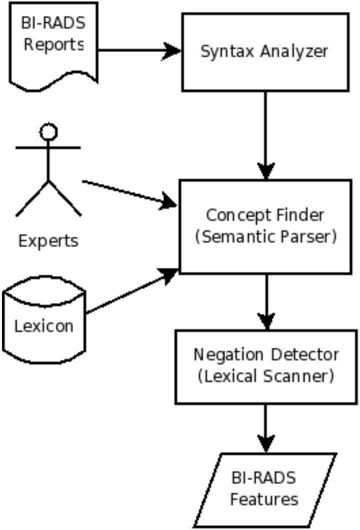


Information from Lexicon

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

Text	Concept
indistinct margin	indistinct margin
indistinct calcification	amorphous calcification
indistinct image	not a BI-RADS concept

Algorithm Flowchart



- 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

NLP Bibliography

- 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.
- **H. Nassif**, F. Cunha, I.C. Moreira, R. Cruz-Correia, E. Sousa, D. Page, E. Burnside, and I. Dutra.
  Extracting BI-RADS Features from Portuguese Clinical Texts.
  BIBM, Philadelphia, 2012. Accepted.
- **H. Nassif**, R. Wood, E.S. Burnside, M. Ayvaci, J. Shavlik and D. Page.
  Information Extraction for Clinical Data Mining: A Mammography Case Study.
  ICDMW, Miami, pp. 37-42, 2009.

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 *

Other Work

- SAYU
- Mammography upgrades
  - I. Dutra, **H. Nassif**, D. Page, J. Shavlik, R. Strigel, Y. Wu, E.M. Elezabi, and E. Burnside.
    Integrating Machine Learning and Physician Knowledge to Improve the Accuracy of Breast Biopsy.
    AMIA, Washington, DC, pp. 349-355, 2011.
- 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

Covalent bond

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

Definition

Electronegativity: Measure of atom's attraction for electrons

Covalent Bonds

$O^{-\delta}$ — $H^{+\delta}$

Covalent bond

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

Definition

Electronegativity: Measure of atom's attraction for electrons

Covalent Bonds

$O^{-\delta} — H^{+\delta}$

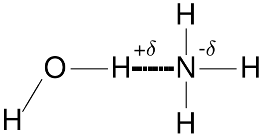
Covalent polar bond

- 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



Hydrogen bond

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

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

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

Outline

7. Appendix A: Hexoses
   - Atomic Interactions
   - Hexose Features
8. Appendix B: Algorithms
   - RF-SVM
   - ILP
   - Instance Relabeling

Hexose Features

Atomic Feature	Values
Charge	Negative, Neutral, Positive
Hydrogen-bonding	Non-hydrogen bonding, Hydrogen-bonding
Hydrophobicity	Hydrophilic, Hydroneutral, Hydrophobic

Residue Grouping	Amino Acids
Aromatic	HIS, PHE, TRP, TYR
Aliphatic	ALA, ILE, LEU, MET, VAL
Neutral	ASN, CYS, GLN, GLY, PRO, SER, THR
Acidic	ASP, GLU
Basic	ARG, LYS

Atomic Chemical Properties I

PDB atom symbol	Residues	Partial Charge	Hydro-phobicity	Hydrogen Bonding
Amino acid oxygen atoms				
O	All amino acids	0	HPHIL	HB
OXT	All amino acids	-ve	HPHIL	HB
OE1, OE2, OD1, OD2	GLU, ASP	-ve	HPHIL	HB
OE1, OD1	GLN, ASN	0	HPHIL	HB
OG, OG1, OH	SER, THR, TYR	0	HPHIL	HB
Amino acid carbon atoms				
C	All amino acids	0	HNEUT	NHB
CA	All amino acids	0	HNEUT	NHB
CB, CG, CD, CE	ALA, SER, THR, CYS, ASP, ASN, GLU, GLN, ARG, LYS, PRO	0	HNEUT	NHB
CB, CG, CD, CE	LEU, VAL, ILE, MET	0	HPHOB	NHB
CG1, CG2, CD1, CD2, CD1	LEU, VAL, ILE	0	HPHOB	NHB
CG, CD1, CD2, CE1, CE2, CZ, CG,CD1, CD2, CE2, CE3, CZ2, CZ3, CH2	PHE, TYR, TRP	0	HPHOB	NHB
CG, CD2, CE1	HIS	0	HPHOB	NHB

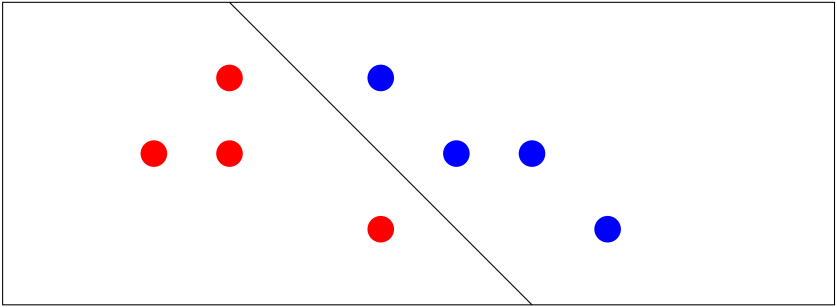
Atomic Chemical Properties II

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

Outline

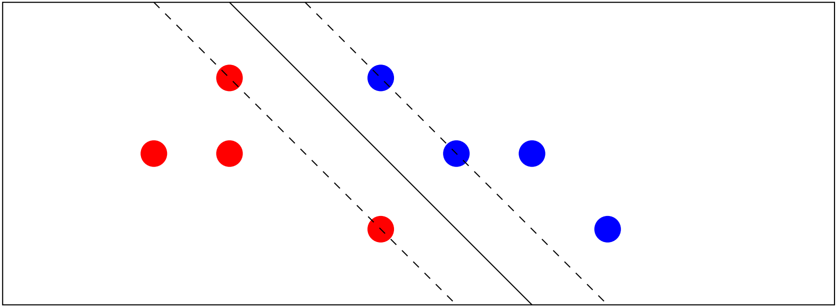
Support Vector Machines (SVM, *Vapnik'98*)

- 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



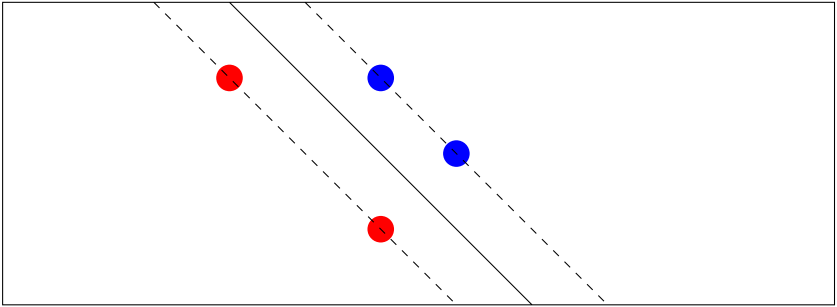
Support Vector Machines (SVM, *Vapnik'98*)

- 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



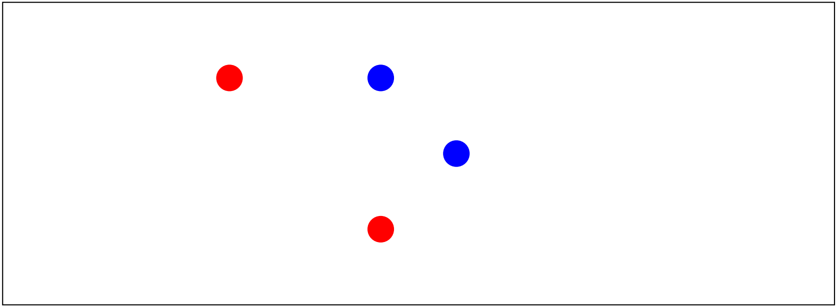
Support Vector Machines (SVM, *Vapnik'98*)

- 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

Outline

7 Appendix A: Hexoses
- Atomic Interactions
- Hexose Features

8 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

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*

ProGolem (Bottom-Up, *Muggleton'09*)

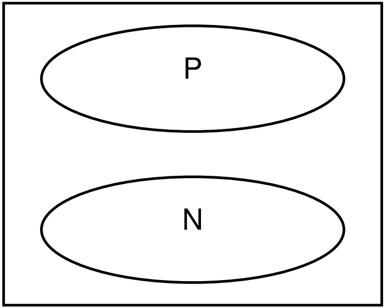
Require: Examples E, mode declarations M, background knowledge B, Scoring function S

$Learned\_rules \leftarrow \{\}$
$Pos \leftarrow$ 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 $Learned\_rules$

Outline

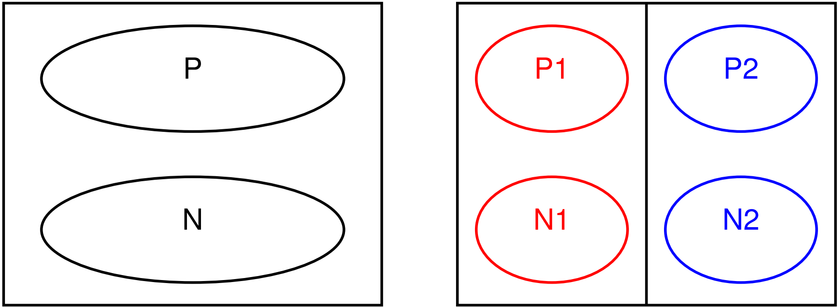
Coverage Scoring Function

- Rule coverage score: $Cover(P) - Cover(N)$
- DP: $(Cover(P1) - Cover(N1)) - (Cover(P2) - Cover(N2))$

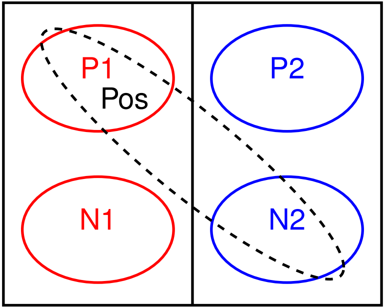


Coverage Scoring Function

- Rule coverage score: $Cover(P) - Cover(N)$
- DP: $(Cover(P1) - Cover(N1)) - (Cover(P2) - Cover(N2))$

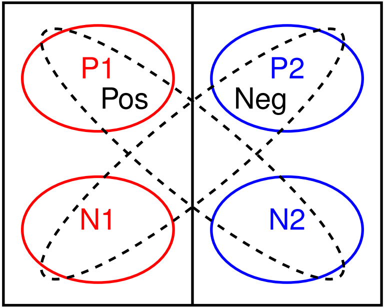


Instance Relabeling DP Method (*Page'12*)



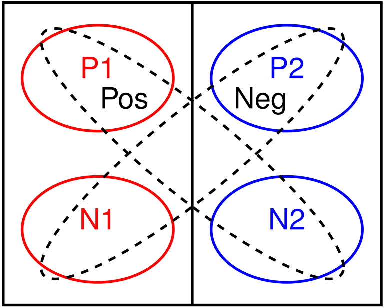
- Relabel $Pos = P1 + N2$
- Relabel $Neg = P2 + N1$
- Run standard ILP
- $Cover(Pos) - Cover(Neg)$
- $Cover(P1+N2) - Cover(P2+N1)$
- $(Cover(P1) + Cover(N2)) - (Cover(P2) + Cover(N1))$
- $(Cover(P1) - Cover(N1)) - (Cover(P2) - Cover(N2))$

Instance Relabeling DP Method (*Page'12*)



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

Instance Relabeling DP Method (*Page'12*)



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