Feature Selection and Discovery
Using Local Structure Induction:
Algorithms, Applications, and Challenges

Constantin F. Aliferis, M.D., P.h.D
Assistant Professor, Department of Biomedical Informatics
Vanderbilt University Medical Center

Stanford Medical Informatics Colloquium
Thursday, April 8th, 2004
Acknowledgments

- Collaborators
  - Ioannis Tsamardinos Ph.D.
  - Douglas Hardin Ph.D.

- Students
  - Alexander Statnikov M.S.
  - Laura Brown M.S.
  - Yindalon Aphinyanaphongs MS

- Support
  - NIH
  - Vanderbilt University
Outline

• **Part I**: Link between the feature selection for classification and local (causal) structure discovery problem

• **Part II**: Novel Algorithms to discover local structure and perform optimal feature selection

• **Part III**: Going from local to global structure discovery

• **Part IV**: Challenges, limitations, open problems
Preliminaries

• Definitions of the problems we will consider
• Why they are important and still open
• What are the basic tools we will use to attempt solutions
• Two primary desiderata:
  – Principled Methods = well characterized, i.e., we know under what conditions they provide a solution to the problem, how good is the solution, when they may fail
  – Scaleable Methods = able to cope with datasets with thousands or even millions of variables such as the ones that started appearing biological sciences, medicine, and a variety of other fields such as business, economics, sociology, military operations. Very often the available data is very high-dimensional but has small sample rendering more traditional statistical approaches impractical or infeasible
Definition: The problem of Feature Selection for Classification

- Select among a set of variables the smallest subset that maximizes performance in some diagnostic, predictive, or other classification task*

* Note: typically for a given set of classifiers and classification performance metrics
Importance of Feature Selection for Classification

• In Biomedicine, Diagnosis, Prognosis, Treatment Selection, and (aspects of) Prevention can often be cast as Classification

• Feature Selection for Classification in that context:
  – May lead to more accurate classifiers
  – Classification algorithms may not scale up to the size of the full feature set either in sample or time
  – Presumably it allows us to better understand the domain
  – Cheaper to collect a reduced set of predictors
  – Safer to collect a reduced set of predictors
…But…Isn’t the Feature Selection Problem Solved?

• Immense variety of algorithms in existence!

• Major types:
  – “Filters” (do not employ the learner of choice);
  – “Wrappers” (employ the learner of choice with heuristic search);
  – Other: Embedded in the learner’s operation

• Feature selection algorithms rarely, if ever, provide guarantees for correctness (either in the theoretical or practical sense)

• Few algorithms scale up to the massive datasets of today
Definition: The Causal Discovery Problem

- Discover the precise causal relationships among the variables in some domain

- Several Variants. We will focus on:
  - Discovery of all relationships (Global Causal Discovery)
  - Discovery of all relationships directly to and from a variable of interest (Local Causal Discovery)
  - Observational i.i.d. data
Isn’t The Causal Discovery Problem Solved With Experiments?

- No! Consider Randomized Controlled Trials (RCTs) for example:
  - Unethical (e.g., a RCT about the effects of smoking)
  - Costly/Time consuming (gene manipulation, epidemiology)
  - Impossible (astronomy)
  - Potentially extremely large number is needed to derive precise structure involving several variables
…Yet Causal Questions Abound

- What SNP combination causes what disease?
- How genes and proteins are organized in complex causal regulatory networks?
- How behaviour causes disease?
- How genotype causes differences in response to treatment?
- How the environment modifies or even supersedes the normal causal function of genes?
- Etc., etc.
An Imperfect Solution: Causal Induction Heuristics

1. Surgeon’s General’s “Epidemiological Criteria for Causality” [Surgeon General of the United States 1964]: $A$ is causing $B$ with high likelihood if:
   - $A$ precedes $B$;
   - $A$ is strongly associated with $B$;
   - $A$ is consistently associated with $B$ in a variety of research studies, populations, and settings;
   - $A$ is the only available explanation for $B$ (“coherence”);
   - $A$ is specifically associated with $B$ (but with few other factors).

2. ‘If $A$ is a robust and strong predictor of $T$ then $A$ is likely a cause of $T$’

3. ‘The closer $A$ and $T$ are in a causal sense, the stronger their correlation’

4. ‘If they cluster together they have similar or related function’
An Imperfect Solution: Causal Induction Heuristics

- Using such heuristics is very common in Epidemiology, Bioinformatics, Biomarker Discovery, and in a variety of non-biomedical fields
An Empirical Demonstration Of How Causal Discovery Heuristics May Fail: Use Generic Feature Selection For Local Causal Discovery

Network: Alarm-1k

(999 variables, consists of 37 tiles of Alarm network)

Training sample size = 1000
Testing sample size = 1000
Target = 407
Target = 407
Feature Selection Method = RFE Linear

The algorithm returned 230 variables

Classification Perf. = 84.29% AUC
Target = 407
Feature Selection Method = RFE Polynomial

The algorithm returned 51 variables

Classification Perf. = 91.79% AUC
Target = 407
Feature Selection Method = BFW

The algorithm returned 36 variables

Classification Perf. = 92.13% AUC
Is This A General Phenomenon
Or A Contrived Example?
# Random Targets in Tiled ALARM

## Evaluation of classification performance

Metric = AUC (%), Classifier = RBF SVM

<table>
<thead>
<tr>
<th>FS method</th>
<th>Alarm-1k</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>936</td>
</tr>
<tr>
<td>HITON MB</td>
<td>65.2</td>
</tr>
<tr>
<td>MMMB</td>
<td>65.2</td>
</tr>
<tr>
<td>RFE Linear</td>
<td>92.2</td>
</tr>
<tr>
<td>RFE Poly</td>
<td>91.0</td>
</tr>
<tr>
<td>BFW</td>
<td>96.3</td>
</tr>
<tr>
<td>No F/S</td>
<td>94.8</td>
</tr>
</tbody>
</table>

## Metric = Proximity score

<table>
<thead>
<tr>
<th>FS method</th>
<th>Alarm-1k</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>936</td>
</tr>
<tr>
<td>HITON MB</td>
<td>10.7</td>
</tr>
<tr>
<td>MMMB</td>
<td>10.7</td>
</tr>
<tr>
<td>RFE Linear</td>
<td>16.0</td>
</tr>
<tr>
<td>RFE Poly</td>
<td>15.4</td>
</tr>
<tr>
<td>BFW</td>
<td>13.5</td>
</tr>
<tr>
<td>No F/S</td>
<td>16.1</td>
</tr>
</tbody>
</table>
# Random Targets in GENE

## Evaluation of classification performance

<table>
<thead>
<tr>
<th>Metric = Accuracy (%)</th>
<th>Classifier = Decision Trees (See 5.0)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>FS method</strong></td>
<td><strong>Gene</strong></td>
</tr>
<tr>
<td></td>
<td>726</td>
</tr>
<tr>
<td>HITON MB</td>
<td>87.5</td>
</tr>
<tr>
<td>BFW</td>
<td>85.1</td>
</tr>
<tr>
<td>No F/S</td>
<td>84.8</td>
</tr>
</tbody>
</table>

## Metric = Proximity Score

<table>
<thead>
<tr>
<th><strong>FS method</strong></th>
<th><strong>Gene</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>726</td>
</tr>
<tr>
<td>HITON MB</td>
<td>0</td>
</tr>
<tr>
<td>MMB</td>
<td>0.333333</td>
</tr>
<tr>
<td>BFW</td>
<td>3.32395</td>
</tr>
<tr>
<td>No F/S</td>
<td>4.0625</td>
</tr>
</tbody>
</table>
Can This Problem Be Solved Effectively?
Target = 407, Method = HITON_MB

Classification Perf. = 90.04% AUC
How Do We Obtain Algorithms That Can Solve The Feature Selection And Discovery Problems? How Are They Related (If At All)?
First Tool: Bayesian Networks (Provide A Formal Language to Model Causality)

- **BN =**
  - Graph (Variables (nodes), dependencies (arcs)) +
  - Joint Probability Distribution +
  - Markov Property
- Graph has to be DAG (directed acyclic) in the **standard** BN model (but extensions do exist)

\[
\begin{align*}
\text{JPD} & \\
P(A+, B+, C+) &= 0.006 \\
P(A+, B+, C-) &= 0.014 \\
P(A+, B-, C+) &= 0.054 \\
P(A+, B-, C-) &= 0.126 \\
P(A-, B+, C+) &= 0.240 \\
P(A-, B+, C-) &= 0.160 \\
P(A-, B-, C+) &= 0.240 \\
P(A-, B-, C-) &= 0.160
\end{align*}
\]
First Tool: Bayesian Networks (Provide A Formal Language to Model Causality)

- **Markov Property**: the probability distribution of any node $N$ given its parents $P$ is independent of any subset of the non-descendent nodes $W$ of $N$

$$
\begin{align*}
D & \perp \{ B, C, E, F, G \mid A \\
F & \perp \{ A, D, E, F, G, H, I, J \mid B, C \}
\end{align*}
$$
First Tool: Bayesian Networks (Provide A Formal Language to Model Causality)

- The Markov property enables us to decompose (factor) the joint probability distribution into a product of prior and conditional probability distributions.

\[ P(V) = \prod_i p(V_i | Pa(V_i)) \]

The original JPD:

- \( P(A^+, B^+, C^+) = 0.006 \)
- \( P(A^+, B^+, C^-) = 0.014 \)
- \( P(A^+, B^-, C^+) = 0.054 \)
- \( P(A^+, B^-, C^-) = 0.126 \)
- \( P(A^-, B^+, C^+) = 0.240 \)
- \( P(A^-, B^+, C^-) = 0.160 \)
- \( P(A^-, B^-, C^+) = 0.240 \)
- \( P(A^-, B^-, C^-) = 0.160 \)

Becomes:

- \( P(A^+) = 0.8 \)
- \( P(B^+ | A^+) = 0.1 \)
- \( P(B^+ | A^-) = 0.5 \)
- \( P(C^+ | A^+) = 0.3 \)
- \( P(C^+ | A^-) = 0.6 \)

Up to Exponential Saving in Number of Parameters
Second Tool: Faithfulness

• Fundamental condition for causal discovery

• Informally a distribution $J$ is faithful to some BN $X$ iff for arbitrary non-overlapping node subsets $A, B, C$ in $X$, $A$ is dependent of $B$ given $C$ whenever the Markov Condition does not imply that they are independent (otherwise they are independent). Hence $X$ is a perfect dependence/independence map of the distribution

• Faithful distributions constitute the vast majority of theoretical distributions (Meek 1995)
Third Tool: A Fundamental Theorem of Computational Causal Discovery (Spirtes, Glymour, Scheines, 1990)

If JPD J is faithful to some BN B:

- An edge $X \rightarrow Y$ (of unknown direction) exists, if and only if for all sets of nodes $S$, $\text{Dep}(X, Y \mid S)$ (allows discovery of the edges)
- If structure $\begin{array}{c}
  B \\
  F \\
  C 
\end{array}$ and for every set $S$ that contains $F$, $\text{Dep}(X, Y \mid S)$, then $\begin{array}{c}
  F \\
  B \\
  C 
\end{array}$

Note: When using search-and-score methods for causal discovery, Faithfulness is implied by the heuristic search needed to render the algorithms tractable.
Third Tool: Relevancy

- (Kohavi and John 1997):
  - Strongly relevant features = always needed to obtain optimal classification (not independent of T given all remaining variables)
  - Weakly relevant features = contain information about the classification but are not necessarily required to achieve optimal classification (not SR and independent of T given some subset of remaining variables)
  - Irrelevant features = do not carry any information about the classification (not SR or WR and independent of T given all subsets of remaining variables)
  - Relevant Features = Strongly or Weakly relevant ones
Part I:
Some Theoretical Results Regarding Feature Selection and its Link to Causal Discovery
Result 1, Corollary 1: Necessary Conditions for Feature Selection Optimality; Link Between Causal Discovery and Feature Selectin

• Given:
  – Faithful distributions
  – Powerful enough learners (with “universal approximator” behavior)
  – Calibrated classification
  – Enough training sample for reliable statistical testing

The Markov Blanket of a target variable T is the minimal set conditioned on which all other variables become independent of T. The Markov Blanket of T is the solution to the feature selection problem. Furthermore the Markov Blanket always exists and is unique.

• Under the above conditions, the solution to the feature selection problem for some variable T:
  – Contains the direct causes and direct effects of T
  – Also contains the direct causes of the direct effects of T
Result 2: Markov Blanket Relevancy Is Equivalent to Previous Notions Of relevancy IN Faithful Distributions

• The Kohavi and John definitions of relevancy can be cast in terms of Bayes Networks in faithful distributions:

  – **Strongly relevant features** = members of the Markov Blanket
  – **Weakly relevant features** = variables with a path to the Markov Blanket but not in the Markov Blanket
  – **Irrelevant features** = variables with no path to the Markov Blanket
Result 3: Non-Superiority of Wrappers Over Filters (and vice versa)

- Neither Wrappers nor Filters are inherently (i.e., over all possible data distributions and metrics) superior over the other approach:
  - Wrappers are subject to the No Free Lunch Theorem for Optimization => some wrappers will be good for specific distributions and metrics and bad for others; there is no wrapper that dominates over any other over all possible distributions/metrics. A random walk in the space of feature subset is on the average as good as any wrapper.
  - There cannot be a definition of relevancy/filter algorithm that does not take into consideration the distribution and metric
Results 4-7: Behavior of SVM-Based Feature Selection

Linear (weight-based) SVM-based feature selection:
- Identifies all Kohavi-John irrelevant features in faithful distributions
- May remove strongly relevant features
- May not remove weakly relevant features
- Does not have a local causal interpretation
Part II: Algorithms to Induce Markov Blankets and Local Causal Structure
**Prior Work**

- **Local Neighborhood (direct causes and direct effects):**
  - No special-purpose prior algorithms! Previous solution: learn the full causal Bayesian Network. However this is not feasible for more than a few hundred variables…

- **Previous Markov Blanket Algorithms:**

| ALGORITHM                  | SOUND | HIGHLY SCALABLE (>10⁴ variables) | SAMPLE EXPONENTIAL TO | MB| (for sound algorithms) | COMMENTS                                                                 |
|----------------------------|-------|----------------------------------|-----------------------|------------------------|--------------------------------------------------------------------------|
| Cheng and Greiner (1999)   | YES   | YES                              |                       |                         | Post-processing on learning BN                                          |
| Cooper et al. (1997)       | NO    | NO                               |                       |                         | Uses full BN learning                                                   |
| Margaritis and Thrun (1999)| YES   | YES                              | YES                   |                        | Intended to facilitate BN learning                                      |
| Koller and Sahami (1996)   | NO    | NO (except for very small k)     |                       |                        | Most widely-cited MB induction algorithm                                |

*Chiara F. Aliferis, Feature Selection and Discovery Using Local Structure Induction*
Desiderata

- We need an algorithm that satisfies *all* of the following:
  - Is sound under well-defined and reasonable conditions
  - Is very highly scaleable to the number of variables
  - Is sample efficient (good starting point: does not require sample that is exponential to the cardinality of the Markov Blanket)
Small-Sample Large Causal Neighborhood (SLCN) algorithm family

• Outline:

1. Use the Fundamental Causal Discovery Theorem to find direct causes and direct effects of the target T.
   To make application of the theorem efficient use an interleaved elimination-admission strategy. Use a good heuristic, i.e., one that considers for admission first variables that are likely to be direct causes/effects and never misses such variables

1. Take the union of the direct causes and direct effects. This set by the properties of the Markov Blanet in Faithful distributions contains the set of direct causes and direct effects of T as well as the set of direct causes of the direct effects (spouses)
• Outline (continued):

3. **Use the Fundamental Causal Discovery Theorem to identify and throw away the false positives**, i.e., direct causes of direct causes, direct causes of spouses, direct effects of direct causes, direct effects or direct effects and direct effects of spouses. Not all such false positives are identifiable by the Fundamental Causal Discovery Theorem in all distributions.

4. **Eliminate any remaining false positives by using a wrapping approach and cross-validation**. This also corrects some deviations from the assumptions of the conditions for the optimality of the Markov Blanket for solving the feature selection problem for classification.
HITON: An instantiation of the SLCN algorithm family geared toward classification

**HITON** (Data $D$; Target $T$; Classifier $A$)

“returns a minimal set of variables required for optimal classification of $T$ using algorithm $A$”

$MB(T) = \text{HITON-MB}(D, T)$ // Identify Markov Blanket

Vars = \text{Wrapper}(MB(T), T, A) // Use wrapping to remove unnecessary variables

Return Vars

**HITON-MB**(Data $D$, Target $T$)

“returns the Markov Blanket of $T$”

$PC = \text{parents and children of } T$ returned by 
\text{HITON-PC}(D, T)$

$PCPC = \text{parents and children of the parents and children or } T$

$CurrentMB = PC \cup PCPC$

// Retain only parents of common children and remove false positives

∀ potential spouse $X$ in $CurrentMB$ and ∀ $Y$ in $PC$:

if not $\exists$ $S$ in $\{Y \cup V - \{T, X\}$ so that $\perp(T; X|S)$

then retain $X$ in $CurrentMB$

else remove it

Return $CurrentMB$

**Wrapper**(Vars, T, A)

“returns a minimal set among variables $Vars$ for predicting $T$ using algorithm $A$ and a wrapping approach”

Select and remove a variable.

If internally cross-validated performance of $A$ remains the same permanently remove the variable.

Continue until all variables are considered.

**HITON-PC**(Data $D$, Target $T$)

“returns parents and children of $T$”

$CurrentPC = {}$

Repeat

Find variable $V_i$ not in $CurrentPC$ that maximizes $\text{association}(V_i, T)$ and admit $V_i$ into $CurrentPC$

If there is a variable $X$ and a subset $S$ of $CurrentPC$ s.t. 

$\perp(X : T|S)$

remove $V_i$ from $CurrentPC$;

mark $V_i$ and do not consider it again

Until no more variables are left to consider

Return $CurrentPC$
**MMMB: An instantiation of the SLCN algorithm family geared toward discovery**

- Same as HITON but:
  - In PC induction subroutine uses instead of a univariate association heuristic the following heuristic: consider the variable with maximum minimal association with the Target where the minimal association is over all subsets of the current set of direct causes/effects.
  - Does not include a wrapper phase
Properties of HITON/MMMB

• Sound for classification if:
  – Data comes from a Faithful distribution
  – Powerful enough learners
  – Enough sample to perform statistical tests of independence reliably

• Sound for local causal discovery if:
  – Data comes from a Faithful distribution
  – Local causal sufficiency (i.e., no confounders are connected to the target) [Note: when this condition is violated, only false positives are introduced]
  – Enough sample to perform statistical tests of independence reliably
Properties of HITON/MMMB (continued)

- They do not require sample that is exponential to the cardinality of the Markov Blanket but sample that ranges from constant up to the cardinality of the Markov Blanket \textit{when it is not possible to use smaller sample}.

- They are very efficient in real-life datasets with more than $10^5$ variables running in a few hours or less on a simple Wintel machine. Of course the (local structure induction) problem per se is intractable in the worst case.

- They are directly parallelizeable.
Experimental Evaluations:

Classification

Benchmark Real Data in 5 Domains
### Table: Dataset Characteristics

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Thrombin</th>
<th>Arrythmia</th>
<th>Ohsumed</th>
<th>Lung Cancer</th>
<th>Prostate Cancer</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Problem Type</strong></td>
<td>Drug Discovery</td>
<td>Clinical Diagnosis</td>
<td>Text Categorization</td>
<td>Gene Expression Diagnosis</td>
<td>Mass-Spec Diagnosis</td>
</tr>
<tr>
<td><strong>Variable #</strong></td>
<td>139,351</td>
<td>279</td>
<td>14,373</td>
<td>12,600</td>
<td>779</td>
</tr>
<tr>
<td><strong>Variable Types</strong></td>
<td>binary</td>
<td>nominal/ordinal/continuous</td>
<td>binary and continuous</td>
<td>continuous</td>
<td>continuous</td>
</tr>
<tr>
<td><strong>Target</strong></td>
<td>binary</td>
<td>nominal</td>
<td>binary</td>
<td>binary</td>
<td>binary</td>
</tr>
<tr>
<td><strong>Sample</strong></td>
<td>2,543</td>
<td>417</td>
<td>2000</td>
<td>160</td>
<td>326</td>
</tr>
<tr>
<td><strong>Vars-to-Sample</strong></td>
<td>54.8</td>
<td>0.67</td>
<td>7.2</td>
<td>60</td>
<td>2.4</td>
</tr>
<tr>
<td><strong>Evaluation metric</strong></td>
<td>ROC AUC</td>
<td>Accuracy</td>
<td>ROC AUC</td>
<td>ROC AUC</td>
<td>ROC AUC</td>
</tr>
<tr>
<td><strong>Design</strong></td>
<td>1-fold c.v.</td>
<td>10-fold c.v.</td>
<td>1-fold c.v.</td>
<td>5-fold c.v.</td>
<td>10-fold c.v.</td>
</tr>
</tbody>
</table>

**Figure 2:** Dataset Characteristics
<table>
<thead>
<tr>
<th>1. Drug Discovery (Thrombin)</th>
<th>UAF*</th>
<th>RFE</th>
<th>HITON</th>
<th>ALL</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>96.12%</td>
<td>93.29%</td>
<td>93.23%</td>
<td>93.69%</td>
</tr>
<tr>
<td>KNN</td>
<td>87.25%</td>
<td>89.71%</td>
<td>92.23%</td>
<td>88.21%</td>
</tr>
<tr>
<td>NN</td>
<td>N/A</td>
<td>92.04%</td>
<td>92.65%</td>
<td>N/A</td>
</tr>
<tr>
<td>Average</td>
<td>91.69%</td>
<td>91.68%</td>
<td>92.7%</td>
<td>90.95%</td>
</tr>
<tr>
<td># of variables</td>
<td>34837</td>
<td>8709</td>
<td>32</td>
<td>139351</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>2. Clinical Diagnosis (Atrial Fibrillation)</th>
<th>UAF*</th>
<th>B/F*</th>
<th>HITON*</th>
<th>ALL*</th>
</tr>
</thead>
<tbody>
<tr>
<td>DTI</td>
<td>73.94%</td>
<td>72.85%</td>
<td>71.87%</td>
<td>73.94%</td>
</tr>
<tr>
<td>KNN</td>
<td>63.22%</td>
<td>63.45%</td>
<td>65.30%</td>
<td>63.22%</td>
</tr>
<tr>
<td>NN</td>
<td>58.29%</td>
<td>60.90%</td>
<td>60.38%</td>
<td>58.29%</td>
</tr>
<tr>
<td>Average</td>
<td>65.15%</td>
<td>65.73%</td>
<td>65.85%</td>
<td>65.15%</td>
</tr>
<tr>
<td># of variables</td>
<td>279</td>
<td>96</td>
<td>63</td>
<td>279</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>3. Text Categorization (OHSUMED)</th>
<th>IG</th>
<th>X*</th>
<th>HITON</th>
<th>ALL*</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>82.43%</td>
<td>85.91%</td>
<td>82.85%</td>
<td>90.50%</td>
</tr>
<tr>
<td>SBCtc</td>
<td>84.18%</td>
<td>86.23%</td>
<td>85.10%</td>
<td>84.25%</td>
</tr>
<tr>
<td>KNN</td>
<td>75.55%</td>
<td>81.76%</td>
<td>80.25%</td>
<td>77.56%</td>
</tr>
<tr>
<td>NN</td>
<td>82.47%</td>
<td>85.27%</td>
<td>83.97%</td>
<td>N/A</td>
</tr>
<tr>
<td>Average</td>
<td>81.16%</td>
<td>84.79%</td>
<td>83.04%</td>
<td>84.10%</td>
</tr>
<tr>
<td># of variables</td>
<td>224</td>
<td>112</td>
<td>34</td>
<td>14373</td>
</tr>
<tr>
<td></td>
<td>UAF*</td>
<td>RFE*</td>
<td>HITON*</td>
<td>ALL*</td>
</tr>
<tr>
<td>----------------------</td>
<td>------</td>
<td>------</td>
<td>--------</td>
<td>------</td>
</tr>
<tr>
<td>SVM</td>
<td>99.32%</td>
<td>98.57%</td>
<td>97.83%</td>
<td>99.07%</td>
</tr>
<tr>
<td>NN</td>
<td>99.63%</td>
<td>98.70%</td>
<td>98.92%</td>
<td>N/A</td>
</tr>
<tr>
<td>KNN</td>
<td>95.57%</td>
<td>91.49%</td>
<td>96.06%</td>
<td>97.59%</td>
</tr>
<tr>
<td>Average</td>
<td>98.17%</td>
<td>96.25%</td>
<td>97.60%</td>
<td><strong>98.33%</strong></td>
</tr>
<tr>
<td># of variables</td>
<td>330</td>
<td>19</td>
<td><strong>16</strong></td>
<td>12,600</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>UAF*</th>
<th>RFE*</th>
<th>HITON*</th>
<th>ALL*</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>98.50%</td>
<td>98.95%</td>
<td>99.10%</td>
<td>99.40%</td>
</tr>
<tr>
<td>NN</td>
<td>98.62%</td>
<td>98.78%</td>
<td>97.95%</td>
<td>99.27%</td>
</tr>
<tr>
<td>KNN</td>
<td>77.52%</td>
<td>86.53%</td>
<td>91.36%</td>
<td>76.94%</td>
</tr>
<tr>
<td>Average</td>
<td>91.55%</td>
<td>94.75%</td>
<td><strong>96.14%</strong></td>
<td>91.87%</td>
</tr>
<tr>
<td># of variables</td>
<td>706</td>
<td>87</td>
<td><strong>16</strong></td>
<td>779</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Av. Over Baseline Algorithms</th>
<th>HITON</th>
<th>ALL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Av. Perf. over classifiers</td>
<td>86.1%</td>
<td><strong>87.1%</strong></td>
<td>86.1%</td>
</tr>
<tr>
<td>Av. variable #</td>
<td>4540</td>
<td><strong>32.3</strong></td>
<td>33,476</td>
</tr>
<tr>
<td>Av. reduction</td>
<td>x 8</td>
<td>x 1124</td>
<td>x 1</td>
</tr>
</tbody>
</table>

**Figure 3:** Task-specific and average model reduction performance (in bold, best performance per row; asterisks indicate that the corresponding algorithm yield the best model or a non-statistically significantly worse model than the best one).
Experimental Evaluations:

Classification

Diagnosis of Stroke (Real Data) - Four Tasks
# Ischemic vs Hemorrhagic Stroke

<table>
<thead>
<tr>
<th></th>
<th>All</th>
<th>HITON with binary discretization</th>
<th>HITON with ternary discretization</th>
<th>Recursive Feature Elimination</th>
<th>Non recursive SVM FS (like UAF)</th>
<th>UAF-KW (evaluation by PSVM)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>No wrapper</td>
<td>SVM wrapper</td>
<td>No wrapper</td>
<td>SVM wrapper</td>
<td>No wrapper</td>
</tr>
<tr>
<td>PSVM</td>
<td>0.7456</td>
<td>0.6909</td>
<td>0.7781</td>
<td>0.7653</td>
<td>0.6612</td>
<td>0.7586</td>
</tr>
<tr>
<td>RSVM</td>
<td>0.8054</td>
<td>0.8126</td>
<td>0.6276</td>
<td>0.8290</td>
<td>0.7676</td>
<td>0.8055</td>
</tr>
<tr>
<td>Number of features in CV; mean (min-max)</td>
<td>29 (29-29)</td>
<td>2.5 (2-5)</td>
<td>2.4 (1-4)</td>
<td>2.9 (2-5)</td>
<td>2.9 (2-4)</td>
<td>2.3 (1-5)</td>
</tr>
<tr>
<td>Number of features in final model</td>
<td>29</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>4</td>
</tr>
</tbody>
</table>
## Stroke vs Non-Stroke

<table>
<thead>
<tr>
<th></th>
<th>All</th>
<th>HITON with ternary discretization</th>
<th>Recursive Feature Elimination</th>
<th>Non-recursive SVM FS (like UAF)</th>
<th>UAF-KW (evaluation by PSVM)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>No wrapper</td>
<td>SVM wrapper</td>
<td>SVM wrapper</td>
<td>Linear</td>
</tr>
<tr>
<td>PSVM</td>
<td>0.8085</td>
<td>0.8020</td>
<td>0.7887</td>
<td>0.7899</td>
<td>0.7915</td>
</tr>
<tr>
<td>RSVM</td>
<td>0.8305</td>
<td>0.8021</td>
<td>0.8122</td>
<td>0.8009</td>
<td>0.8039</td>
</tr>
<tr>
<td>Number of features in CV: mean (min-max)</td>
<td>29 (29-29)</td>
<td>15.7 (12-19)</td>
<td>11.5 (9-14)</td>
<td>12.2 (8-15)</td>
<td>11.6 (9-15)</td>
</tr>
<tr>
<td>Number of features in final model</td>
<td>29</td>
<td>18</td>
<td>17</td>
<td>13</td>
<td>13</td>
</tr>
</tbody>
</table>
## Stroke vs Mimic

<table>
<thead>
<tr>
<th></th>
<th>All</th>
<th>HITON with ternary discretization</th>
<th>Recursive Feature Elimination</th>
<th>Non-recursive SVM FS (like UAF)</th>
<th>UAF-KW (evaluation by PSVM)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PSVM</td>
<td>No wrapper</td>
<td>SVM wrapper</td>
<td>SVM wrapper</td>
<td>Linear</td>
</tr>
<tr>
<td></td>
<td>RSVM</td>
<td>0.8821</td>
<td>0.8729</td>
<td>0.8605</td>
<td>0.8664</td>
</tr>
<tr>
<td>Number of features in CV: mean (min-max)</td>
<td>0.8891</td>
<td>0.8725</td>
<td>0.8743</td>
<td>0.8664</td>
<td>0.8707</td>
</tr>
<tr>
<td>29 (29-29)</td>
<td>8.7 (4-11)</td>
<td>7.3 (4-10)</td>
<td>7.7 (5-11)</td>
<td>12.3 (8-22)</td>
<td>4.2 (10-25)</td>
</tr>
<tr>
<td>Number of features in final model</td>
<td>29</td>
<td>10</td>
<td>6</td>
<td>7</td>
<td>9</td>
</tr>
</tbody>
</table>
Ischemic vs Mimic

<table>
<thead>
<tr>
<th></th>
<th>All</th>
<th>HITON with ternary discretization</th>
<th>Recursive Feature Elimination</th>
<th>Non-recursive SVM FS (like UAF)</th>
<th>UAF-KW (evaluation by PSVM)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No wrapper</td>
<td>SVM wrapper</td>
<td>SVM wrapper</td>
<td>Linear</td>
<td>Polynomial</td>
</tr>
<tr>
<td>PSVM</td>
<td>0.8463</td>
<td><strong>0.8558</strong></td>
<td>0.8268</td>
<td>0.8455</td>
<td>0.8445</td>
</tr>
<tr>
<td>RSVM</td>
<td>0.8379</td>
<td>0.8394</td>
<td><strong>0.8323</strong></td>
<td>0.7931</td>
<td>0.8445</td>
</tr>
<tr>
<td>Number of features in CV: mean (min-max)</td>
<td>29 (29-29)</td>
<td>11.9 (8-14)</td>
<td>9.4 (7-13)</td>
<td>9 (5-11)</td>
<td>17.4 (7-23)</td>
</tr>
<tr>
<td>Number of features in final model</td>
<td>29</td>
<td>14</td>
<td>11</td>
<td>12</td>
<td>16</td>
</tr>
</tbody>
</table>
Experimental Evaluations:

Local Causal Discovery

Simulated Data from Five Benchmark Networks/Distributions
Datasets

- Small networks
  - ALARM, 37 vars
  - Hailfinder, 56 vars
  - Pigs, 441 vars
  - Insurance, 27 vars
  - Win95Pts, 76 vars

- Large networks (tiled versions)
  - ALARM-5K (5000 vars)
  - Hailfinder-5K
  - Pigs-5K

- All variables act as targets in small networks, 10 in large networks
Discovering the Markov Blanket
(Small Networks)

Comparison of MB(T) algorithms on the small BNs

Number of Training Instances

Distance

GS
IAMB
IAMBnPC
KS0
KS1
KS2
MMMB
PC
Discovering the Markov Blanket (Large Networks)

Comparison of MB(T) algorithms on the large BNs

Distance

Number of Training Instances

GS
IAMB
IAMBnPC
MMMB
Discovering the Markov Blanket (Interpretation of Results)

• Distance of 0.1: sensitivity, specificity = 93%
• Distance of 0.2: sensitivity, specificity = 86%
• Average Distance of MMMB with 5000 samples: 0.1
• Average Distance of MMMB with 500 samples 0.2
• Example: distance=0.16, ALARM-5K, 5000 sample size
Part III: From Local to Global Causal Discovery
Algorithm Max-Min Hill Climbing

- **Step #1**: Find candidate parents and children of every variable $V_i$ in $V$; use conditional independence testing to eliminate false positives; call the resulting set $PC(V_i)$
- **Step #2**: Produce an undirected graph such that there is an edge between $X$ and $Y$ if and only if $X$ is in $PC(Y)$ and $Y$ is in $PC(X)$, where $X$, $Y$ are in $V$
- **Step #3**: Orient the edges by applying Bayesian search-and-score (search=hill climbing with operators add edge, remove edge, orient edge, change orientation of edge; score is typically BDeu)
Algorithm Max-Min Hill Climbing (continued)

• MMHC is very similar to Friedman’s Sparse Candidate. The latter’s simple (but powerful) idea is to constrain the search by considering only nodes that are likely to be direct causes when applying add-edge operators to some node.

• Sparse Candidate (SC) is a prominent BN learning algorithm that scales up to hundreds of variables

• MMHC differences from (and anticipated advantages over) SC:
  – MMHC employs a non-heuristic method to produce a superset of the parent set for each node
  – MMHC does not require uniform sparseness in the data-generating BN
  – MMHC does not require that the user provides as input the maximum degree of connectivity
Experimental Evaluations:

Global Causal Discovery

Simulated Data from Five Benchmark Networks/Distributions

Compare Against Three Prototypical Bayesian Network Induction Algorithms (PC, Sparse Candidate, TPDA)

Measures of Comparison:
- BDeu score (log of the probability of the BN given the data)
- Number of structural errors: wrong additions, deletions, or reversal of edges
- KL divergence (measures how well the induced BN captures the data distribution)
- Execution time
### MMHC versus Sparse Candidate, TPDA, PC: Ability To Capture The True Distribution

Table 1: KL Statistic Results. A larger KL implies closer approximation of the distribution. Abs columns report the actual KL statistic for MMHC and all other columns the normalized statistic by MMHC’s KL (e.g., a ratio greater than 1 implies a better performance than MMHC).

<table>
<thead>
<tr>
<th></th>
<th>Sample Size 500</th>
<th>Sample Size 5000</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MM HC</td>
<td>PC</td>
</tr>
<tr>
<td>Abs Ratio with MMHC’s KL</td>
<td>Abs Ratio with MMHC’s KL</td>
<td></td>
</tr>
<tr>
<td>Alarm1</td>
<td>17</td>
<td>0.64</td>
</tr>
<tr>
<td>Alarm3</td>
<td>32</td>
<td>0.66</td>
</tr>
<tr>
<td>Alarm5</td>
<td>56</td>
<td>–</td>
</tr>
<tr>
<td>Alarm10</td>
<td>110</td>
<td>–</td>
</tr>
<tr>
<td>Child</td>
<td>7</td>
<td>0.70</td>
</tr>
<tr>
<td>HailFinder</td>
<td>28</td>
<td>–</td>
</tr>
<tr>
<td>Munin</td>
<td>81</td>
<td>–</td>
</tr>
<tr>
<td>Pigs</td>
<td>179</td>
<td>–</td>
</tr>
<tr>
<td>Gene</td>
<td>276</td>
<td>0.88</td>
</tr>
<tr>
<td>Average</td>
<td>0.72</td>
<td>0.72</td>
</tr>
</tbody>
</table>

C.F. Aliferis, Feature Selection and Discovery Using Local Structure Induction
# MMHC versus Sparse Candidate, TPDA, PC: Ability To Capture The True Structure

Table 2: Structural Hamming Distance (SHD) Results. A lower SHD implies less structural errors made. Abs columns report the actual SHD for MMHC and all other columns the normalized statistic by MMHC’s SHD (e.g., a ratio greater than 1 implies a worse performance than MMHC).

<table>
<thead>
<tr>
<th>Sample Size 500</th>
<th>Sample Size 5000</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>MM HC</strong></td>
<td><strong>PC</strong></td>
</tr>
<tr>
<td><strong>Abs</strong></td>
<td><strong>Ratio with MMHC’s SHD</strong></td>
</tr>
<tr>
<td>Alarm1</td>
<td>29</td>
</tr>
<tr>
<td>Alarm3</td>
<td>107</td>
</tr>
<tr>
<td>Alarm5</td>
<td>220</td>
</tr>
<tr>
<td>Alarm10</td>
<td>546</td>
</tr>
<tr>
<td>Child</td>
<td>16</td>
</tr>
<tr>
<td>HailFinder</td>
<td>178</td>
</tr>
<tr>
<td>Mumin</td>
<td>532</td>
</tr>
<tr>
<td>Pigs</td>
<td>153</td>
</tr>
<tr>
<td>Gene</td>
<td>314</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td><strong>2.83</strong></td>
</tr>
</tbody>
</table>

C.F. Aliferis, Feature Selection and Discovery Using Local Structure Induction
### MMHC versus Sparse Candidate, TPDA, PC: Efficiency

Table 3: Time Results. *Abs* columns report the actual *Time* in seconds for MMHC and all other columns the normalized statistic by MMHC’s Time (e.g., a ratio greater than 1 implies MMHC was faster). *SC* becomes relatively slower than MMHC as the number of variables increases in Alarm1–Alarm10.

<table>
<thead>
<tr>
<th></th>
<th>Sample Size 500</th>
<th>Sample Size 5000</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Abs</td>
<td>Ratio with MMHC’s Time</td>
</tr>
<tr>
<td></td>
<td>MM HC</td>
<td>PC</td>
</tr>
<tr>
<td>Alarm1</td>
<td>6</td>
<td>1.37</td>
</tr>
<tr>
<td>Alarm3</td>
<td>33</td>
<td>1.26</td>
</tr>
<tr>
<td>Alarm5</td>
<td>85</td>
<td>100.07</td>
</tr>
<tr>
<td>Alarm10</td>
<td>391</td>
<td>401.50</td>
</tr>
<tr>
<td>Child</td>
<td>3</td>
<td>0.55</td>
</tr>
<tr>
<td>HailFinder</td>
<td>13</td>
<td>1.54</td>
</tr>
<tr>
<td>Munin</td>
<td>351</td>
<td>–</td>
</tr>
<tr>
<td>Pigs</td>
<td>1888</td>
<td>0.42</td>
</tr>
<tr>
<td>Gene</td>
<td>4966</td>
<td>1.15</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td><strong>63.48</strong></td>
<td><strong>45.00</strong></td>
</tr>
</tbody>
</table>
Figure 5: Performance of the Sparse Candidate as $k$ increases on reconstructing Alarm from 5000 cases.
**MMHC versus Sparse Candidate: Scaleability**

![Graph showing the comparison between MMHC and Sparse Candidate for 5000 sample size](image_url)
An Explanation Of MMHC’s Results

- Errors do not propagate in MMHC. In standard global algorithms local structural errors propagate → MMHC has only one source of structural errors while the baseline algorithms two.
- Different order in executing conditional tests of independence → faster than PC, TPDA
- No need to provide maximum number of parents as input (if it is wrong the quality of the induced BN may be severely affected due to both local errors and errors that propagate); more powerful identification of likely parents set → less structural errors than SC
- Since a good estimate of the parent set for each node is achieved early on, algorithm is faster (i.e., avoids iterations) → faster than SC
Part IVa: Other Potential Applications of Markov Blanket Induction
C.F. Aliferis, Feature Selection and Discovery Using Local Structure Induction

Distributed Learning
Data Compression

• E.g., instead of storing all data related to a patient, store only Markov Blanket information for a set of target variables of choice
Control For Hidden Bias

- Task: Classify Medline Records to content and methodological quality criteria according to the ACPJ EBM criteria and corpus using a machine learning approach. The best classifiers in this task achieve near-perfect discriminatory performance (Aphinyanaphongs and Aliferis in AMIA 2003).
- Open question: are the criteria in the ACPJ corpus the same as the explicit criteria? (or the editors exhibit biases in selecting the best articles for publication)
- Approach: if Markov Blanket of the target (content-quality) does not include non-methodological terms in the journals Purpose then there is no bias (Aphinyanaphongs and Aliferis to appear in Medinfo 2004)
Convert Black Box Models To Models That Can Be Explained & Can Be Used By Boolean Search Engines

• Poly-SVMs in the previous example perform very well but are not easy to run without a specialized system.
• Approach:
  – select features using a Markov Blanket algorithm,
  – build decision trees using only the Markov Blanket,
  – convert trees into boolean queries that are run sequentially
• This renders manageable queries that outperform regular (embedded) decision tree feature selection, does not include non-methodological terms, and have classification accuracy close to the SVM best model (Aphinyanaphongs and Aliferis to appear in Medinfo 2004)
Part IVb: Challenges & Open Problems
Non-Faithful Distributions

• Deterministic functions
• Epistatic (a.k.a. “highly non-linear”) functions
• Cell aggregation in genomic/proteomic assays
• Temporal aggregation in sparsely-sampled time series
Non-Uniform Cost Feature Selection

- Some variables have different costs to observe than others
- Some variables may come in “packets” or “panels” with one cost associated with all variables in the group
“Manipulation” Markov Blanket

- Minimal set of (manipulatable) variables needed to \textit{maximally manipulate} target T
- Minimal (non-uniform) cost set of (manipulatable) variables needed to \textit{maximally manipulate} target T
Modeling Biological Systems

• Causal Bayesian Network is crude for such modeling needs. Probably needs to be used as filtering device and to be combined with more fine-grain mathematical and computational formalisms
• Some biological systems are very highly interconnected (e.g., CNS)
• Feedback loops ➔ false positives
• Dynamical systems in equilibria and outside equilibria
• Tiny samples
• What does the Markov Blanket of a disease $D$ mean?
Other Types of Computational Causal Discovery Questions

- Discover some relationships (Partial Causal Discovery)
- Use mixtures of experimental and observational data
- Interplay between computational causal theory induction and design and execution of experiments
- Mix different types of experimental designs
Goal is to find full network of (causal) relationships
Goal is to find causal order
Goal is to find Markov Blanket
(=> optimal prediction,
And tight superset of direct causes and direct effects)
Goal is to find direct causes
Goal is to find direct effects
Goal is to find
All upstream direct causes
Goal is to find all Downstream direct effects
Goal is to find “chokepoints” suitable of being drug targets.
Causal Insufficiency

- More benign problem in local discovery (creates only false positives)...
- May be addressed by use of the FCI algorithm
Very Small Samples

• Currently Constrained-based methods rely on heuristics for when to conduct a statistical test of independence and when to rely on a default (e.g., assume dependence)

• How can these heuristics be formalized? What are “smart” defaults based on the nature of distribution, available sample and other data properties?

• Are there sample-efficient approximations of conditional independence tests?
Exploit Constrained Distributions

- Monotone Faithfulness
- Deterministic functions
- Domain-specific functions
Exploit Experimental Designs

- Case-control data
- Time Series
- Cell line experiments
- Mixes of observational and experimental data
- Iterative collection and analysis of data
- Incorporation of prior domain knowledge as constraints or priors
Problems With Validation Of Feature Selection Methods

- Difficult, if not impossible, to know truly optimal set, thus evaluation is always relative to other algorithms and not absolute
- Clean experiments can be conducted when data is generated by a known faithful BN or other generating model with clear Markov Blanket
Problems With Validation Of Causal Discovery Methods

• Experimental Design 1:
  – Known structure $\rightarrow$ simulated data $\rightarrow$ recovered structure
  – Problems: limited set of available structures; they may not correspond to distributions of interest

• Experimental Design 2:
  – Real Data $\rightarrow$ evaluation by expert or prior literature
  – Problems: expert may be wrong; existing literature may be incomplete or based on methods with different inductive bias

• Experimental Design 3:
  – Real data $\rightarrow$ experimental verification
  – Problems: experiments are difficult, timely, costly, or even impossible to conduct depending on the domain
Discovery Systems Laboratory
For more Information (causal discovery tools, publications, contact information)
http://discover1.mc.vanderbilt.edu/discover/public/