



Discovering Markov Blankets: Finding Independencies Among Variables

Motivation: Toward Optimal Feature Selection. Koller and Sahami.
Proc. 13th ICML, 1996.

Algorithm: Algorithms for Large Scale Markov Blanket Discovery.
Tsamardinos, *et al.* Proc. 16th FLAIRS, 2003.

Applications: HITON, A Novel Markov Blanket Algorithm for Optimal
Variable Selection. Aliferis, *et al.* TR, Vanderbilt University,
DSL-03-08, 2003.

Presented by: Nakul Verma
May 3, 2005.



Outline

- Motivation
- Introduction to Bayesian Networks and Markov Blankets
- Markov Blanket Discovery algorithms
- IAMB algorithm and results
- HITON algorithm and results



Outline

- Motivation
- Introduction to Bayesian Networks and Markov Blankets
- Markov Blanket Discovery algorithms
- IAMB algorithm and results
- HITON algorithm and results



Selecting Optimal Subsets of Features

- Idea: Select the most relevant subset of features, that is, a small subset which still provides a high classification accuracy.
- Feature selection is an effective technique in reducing dimensionality, removing irrelevant data, increasing learning accuracy, and improving comprehensibility.
- Algorithms for feature selection (FS) fall in two categories:
 - Filter Methods
 - Wrapper Methods



Filter Methods

- Filter methods select a subset of features without involving any learning algorithm. Therefore,
 - FS is a preprocessing step before induction.
 - FS algorithm and learning algorithm don't interact.
 - Filter methods don't inherit any bias of the learning algorithms.
- Example: FOCUS algorithm (exhaustive search on all feature subsets) [Almuallim & Dietterich 1991]



Wrapper Methods

- The Wrapper methods search through the space of feature subsets using the estimated accuracy from an induction algorithm as the measure of goodness for a particular subset of features. Therefore,
 - Algorithms using wrapper methods tend to be computationally more expensive compared to their 'filter' counterparts.
 - A predetermined learning algorithm is needed to measure performance of Wrapper algorithms.
 - Algorithms using the wrapper method tend to give a better performance compared to those using filter methods.
- Example: HITON algorithm (discussed in detailed later)



Using Feature Selection for Good Classification

- Let \mathbf{F} be the complete feature vector taking values \mathbf{f} for one example, and C be the class random variable taking values c , then

$$\Pr(C = c \mid \mathbf{F} = \mathbf{f})$$

is the probability that the class is c , given that the feature values are \mathbf{f} .

- Now considering the reduced feature space,
Let \mathbf{G} be a subset of \mathbf{F} taking values \mathbf{f}_G (projection of \mathbf{f} onto \mathbf{G}).
We want to choose G such that

$$\Pr(C = c \mid \mathbf{G} = \mathbf{f}_G)$$

is as close to $\Pr(C = c \mid \mathbf{F} = \mathbf{f})$ as possible.



Information-Theoretic Measure of Closeness of Distributions

- Let μ and σ be two distributions over some probability space Ω . Then cross-entropy from μ to σ is defined as:

$$D(\mu, \sigma) = \sum_{x \in \Omega} \mu(x) \log \frac{\mu(x)}{\sigma(x)}$$

- This is also known as the Kullback-Leibler (KL) distance. Intuitively, it is a distance function from a "true" probability distribution, μ , to a "guessed" probability distribution, σ .
- So, in feature subset selection problem, we want
 - \mathbf{f} to be μ ("true")
 - \mathbf{f}_G to be σ ("gussed")



Information-Theoretic Approach (Cont.)

- Thus, we want to find a feature subset \mathbf{G} , such that

$$\Delta_G = \sum_{\mathbf{f}} \Pr(\mathbf{F} = \mathbf{f}) D(\Pr(C | \mathbf{F} = \mathbf{f}), \Pr(C | \mathbf{F}_G = \mathbf{f}_G))$$

is close to zero.

- Note that the computation requires knowledge of conditional distributions (C given \mathbf{F} and C given \mathbf{F}_G)



Difficulties of this Approach

- We only get to observe a small sample of examples, which makes it fairly hard to approximate the true distribution to calculate the relative error.
- It is impractical to compute the error Δ_G . It requires exponential number of computations with respect to number of features in the domain.
- We need an alternative to finding conditional distributions.



Outline

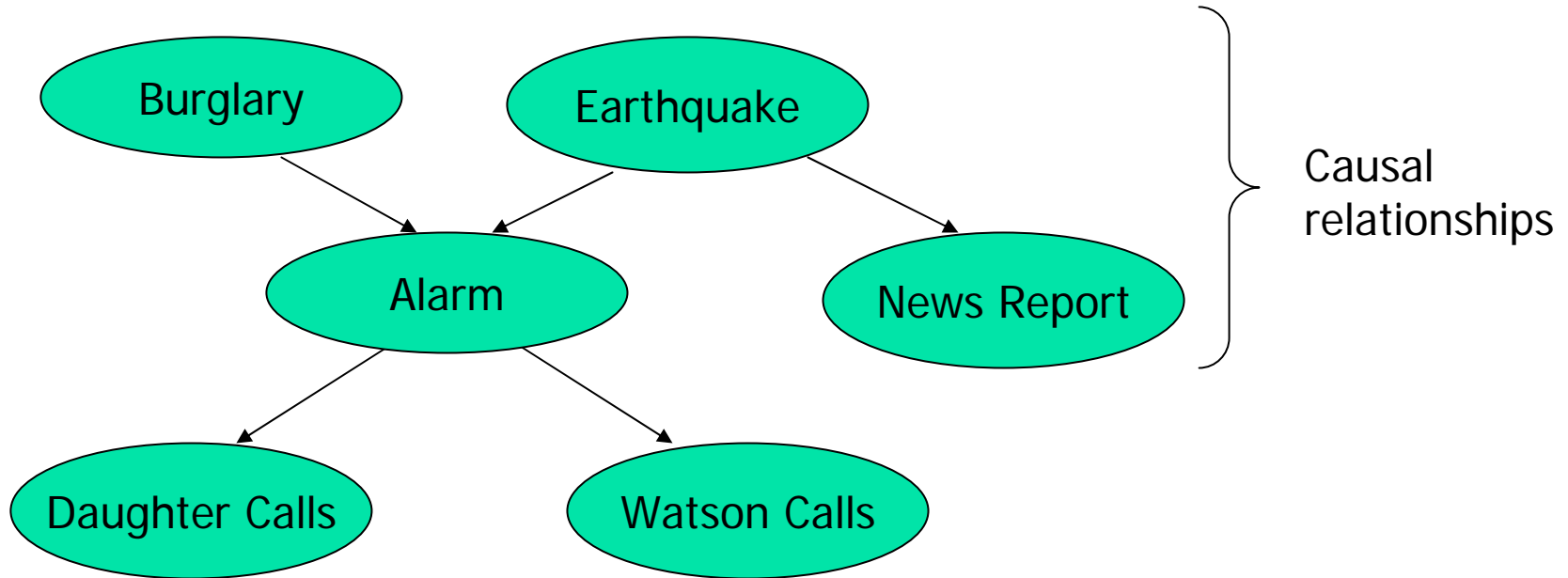
- Motivation
- Introduction to Bayesian Networks and Markov Blankets
- Markov Blanket Discovery algorithms
- IAMB algorithm and results
- HITON algorithm and results



Bayesian Networks (BNs)

- BNs are also known as Belief networks
- A BN is a graph, where nodes are random variables and edges are direct relationships between variables.
- BNs are used for inference: given observations of some nodes, one wants to know the probability distribution of other nodes.
- Inference methods for BN can be classified into two categories: exact reasoning and sampling.

An Example (Pearl '88)



- $\Pr(\text{Burglary} | \text{Alarm}, \text{Report})?$
- $\Pr(\text{Alarm} | \neg \text{Burglary})?$

Need to encode information exponential in the number of parents as Conditional Probability Tables.

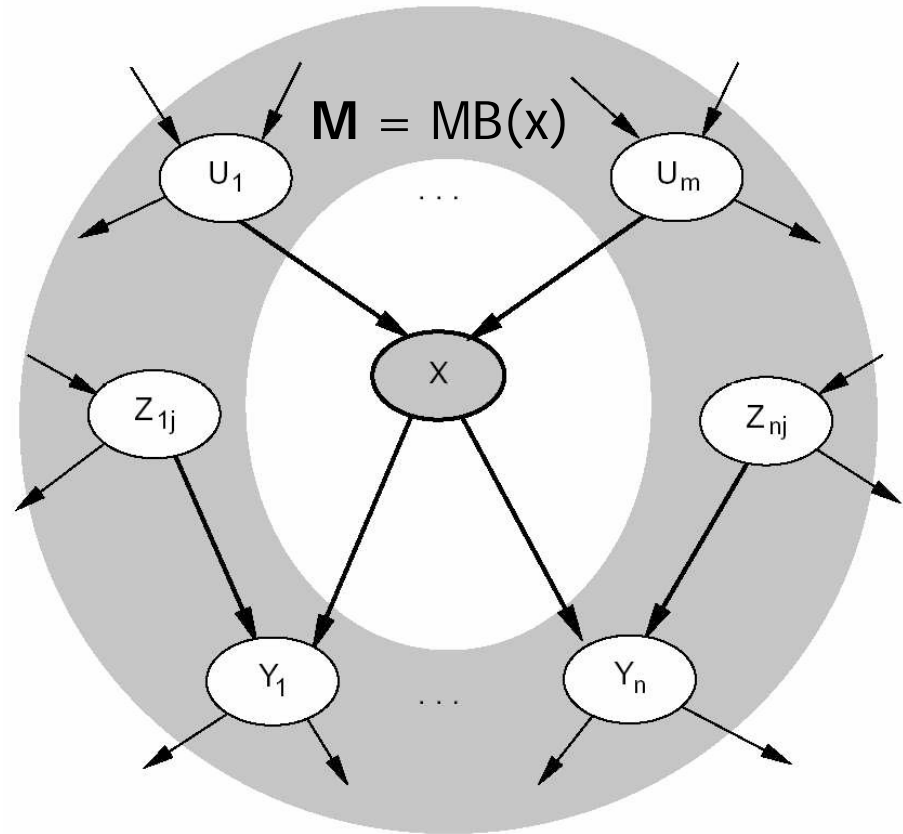


Some Properties of BNs

- A node is independent of its non-descendants given its parents
- A node is independent of all other nodes, given its Markov blanket. So, what is a Markov blanket?

Markov Blanket (MB)

- The Markov Blanket of a node is the set consisting of its parents, children, and spouses.
- More formally:
Let \mathbf{N} be the set of all nodes and \mathbf{M} be a set of nodes not containing x , then \mathbf{M} is a MB for x if x is conditionally independent of $\mathbf{N} - \mathbf{M} - x$ given \mathbf{M} .
 \mathbf{M} is minimal

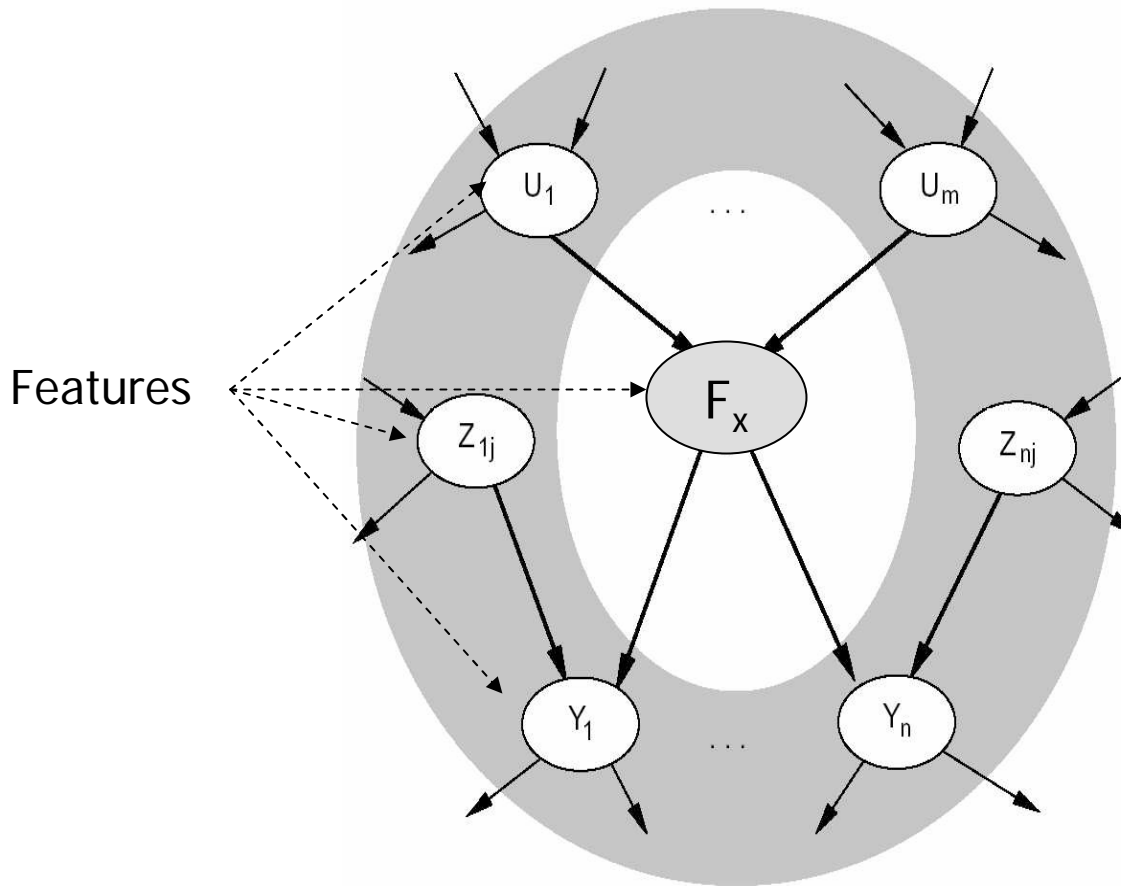




Why are MBs interesting?

- MBs help in studying how an attribute x “behaves” under the effect of other attributes in the domain, by providing ‘shielding’ information.
- MBs can help determine causal relationships among various nodes in a BN
- They can help determine the structure of a BN given just its nodes.
- They can be used in finding good feature subsets.

Connection between MBs and feature selection



Feature F_x is independent of all other features given
 $MB(F_x) = \{U_1 \dots U_m, Z_{1j} \dots Z_{nj}, Y_1 \dots Y_n\}$

F_x gives no extra information regarding the BN given its Markov Blanket.



Using Markov Blankets for Better Feature Selection

- Algorithm idea:
 - If we can find Markov Blanket for a feature F_i , remove feature F_i .
 - Return the remaining features as the minimal set.
- But, if we remove feature F_i based on MB \mathbf{M} , later, we might remove some other feature $F_j \in \mathbf{M}$. Does the removal of F_j may make F_i relevant again?

Theorem: Let \mathbf{G} be the current set of features, and assume that $F_i \notin \mathbf{G}$ has a MB within \mathbf{G} . Let $F_j \in \mathbf{G}$ be some feature which we are about to remove, Then F_i also has a MB within $\mathbf{G} - \{F_j\}$

[Koller & Sahami, '96]



Implication of the theorem

Using Markov blankets for feature elimination has desirable properties:

- We can eliminate a conditionally independent feature F_i , without increasing our distance, Δ_G , from the desired distribution.
- Markov blanket criterion only removes attributes that are really unnecessary: attributes that are irrelevant to the target concept, and attributes that are redundant given other attributes.

[Koller & Sahami, '96]



Outline

- Motivation
- Introduction to Bayesian Networks and Markov Blankets
- Markov Blanket Discovery algorithms
- IAMB algorithm and results
- HITON algorithm and results



Markov Blanket Discovery

Some early approaches:

- KS (Koller-Sahami) algorithm ('96)
 - For each feature F_i in G , let M_i be the set of K features F_j in $G - F_i$ for which expected cross-entropy is the minimum.
 - Compute error, Δ_G of $(F_i|M_i)$ for each i
 - Choose the i for which this quantity is the minimal
- GS (Grow-Shrink) algorithm (Margaritis and Thrun '99)
 - Statically orders the variables according to the strength of association with T .
 - Thus, has some limitations of employing potentially inefficient heuristics.
- PC (Spirtes et al. '00)
 - BN learning algorithm, that is, learns the whole network.
 - Starts as fully connected BN graph, and removes redundant edges until a sound BN remains.
 - MB can then be read off from the resulting network.



Outline

- Motivation
- Introduction to Bayesian Networks and Markov Blankets
- Markov Blanket Discovery algorithms
- **IAMB algorithm and results**
- HITON algorithm and results

IAMB algorithm (Tsamardinos, *et al.*)

Phase I (forward)

$CMB = \emptyset,$

While CMB has changed

Find the feature X in $V - CMB - \{T\}$ that maximizes $f(X; T | CMB)$

If not $I(X; T | CMB)$

Add X to CMB

End If

End While

Phase II (backwards)

Remove from CMB all variables X , for which $I(X; T | CMB - \{X\})$

Return CMB

CMB is the current MB

Heuristic approach for finding Markov blanket of T .

Conditional Independence test.

$$\Pr(X, Y | CMB) = \Pr(X | CMB) \Pr(Y | CMB)$$



IAMB algorithm

Phase I (forward)

$CMB = \emptyset$,

While CMB has changed

Find the feature X in $V -$

$CMB - \{T\}$ that maximizes $f(X; T | CMB)$

If not $I(X; T | CMB)$

Add X to CMB

End If

End While

Phase II (backwards)

Remove from CMB all vari-

ables X , for which $I(X; T$
 $| CMB - \{X\})$

Return CMB

- IAMB is an abbreviation for:
Incremental **A**ssociation **M**arkov **B**lanket
- As we can see, it is a two phase algorithm
 - Growing phase
Adds variables which are part of $MB(T)$ – plus more, i.e., false positives
 - Shrinking phase
Removes false positives.
- Result: Markov blanket for a particular variable T



IAMB algorithm

Heuristic to identify potential Markov Blanket members:

- Include the variable that maximizes a heuristic function $f(X ; T | \text{CMB})$.
- Function f should be non-zero value for every X that is a member of the Markov Blanket of T .
- Typically it is a measure of association between X and T given CMB.
- The authors use the Mutual Information formula for f :
 - $f(X;T | \text{CMB}) = H(X | \text{CMB}) - H(X | T, \text{CMB})$
 - The information T tells us about X is the reduction in uncertainty about X due to the knowledge of T , given the CMB.
 - Computationally, it takes linear time in number of variables



IAMB Variants

Authors also present some variations on the IAMB algorithm.

- InterIAMB
 - It interleaves the growing phase of IAMB (phase I) with the shrinking phase (phase II) attempting to keep the size of MB(T) as small as possible during all steps of the algorithm's execution.

- IAMBnPC
 - It substitutes the shrinking phase (phase II) as implemented in IAMB with the PC algorithm instead

- InterIAMBnPC
 - Combines the two approaches above to reduce the size of the conditioning sets.



Results

Data-sets:

- Real world datasets:
 - ALARM Network (Beinlich, *et al.* '89) – BN used in medical domain, having 37 variables.
 - Hailfinder (Abramson, *et al.* '96) – BN used for modeling weather, with 56 variables.
- Randomly generated BNs:
 - BN with 50 nodes
 - BN with 200 nodes
 - BN with 1000 nodes

} 0-10 Parents chosen randomly for each node



Results

Evaluation metric:

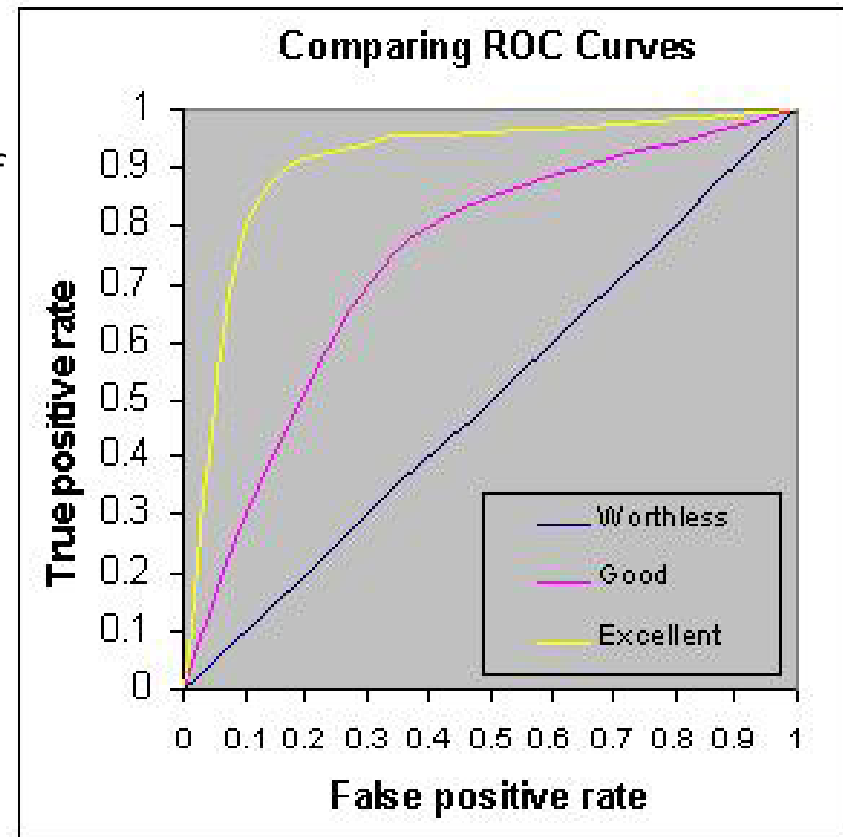
- Area under the ROC Curve (explained next).

Threshold parameters:

- PC: significance levels of G^2 statistical test.
- GS / IAMB Variants: $\text{Mutual-Info}(X;T|\text{CMB}) < \text{threshold}$
- KS: all possible values of the parameter k

Receiver Operating Characteristic (ROC) Curve

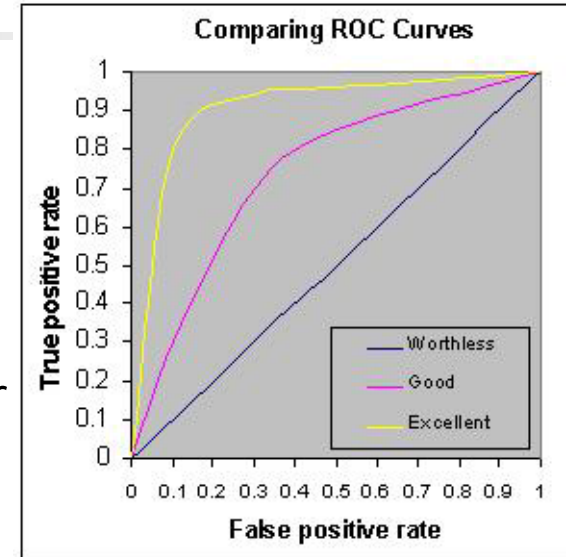
ROC plots true positive rate (TPR) against the false positive rate (FPR) for the different possible thresholds of a diagnostic test



ROC Curve (Cont'd)

A ROC curve demonstrates several things:

- It shows the tradeoff between TPR and FPR.
- The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test.
- The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test.
- The area under the curve (AUC) is a measure of test accuracy. Higher the better.





Results

	<i>ALARM</i>	<i>HAILFINDER</i>				
		<i>Target 1</i>	<i>Target 2</i>	<i>Target 3</i>	<i>Target 4</i>	<i>Average</i>
IAMB	86.70	96.30	96.23	97.12	78.04	90.88
interIAMB	86.70	96.30	96.23	97.12	78.04	90.88
interIAMBnPC	90.50	100.00	100.00	97.12	78.04	93.13
IAMBnPC	89.30	100.00	100.00	97.12	78.04	92.89
GS	80.59	96.30	77.67	72.12	68.04	78.94
KS, k=0	82.82	100.00	92.31	88.73	97.60	92.29
KS, k=1	80.56	70.28	47.76	82.84	67.40	69.77
KS, k=2	82.14	99.53	42.95	45.59	75.00	69.04
PC	95.20	99.07	98.11	81.73	96.08	94.04

Table 1: Experiments on Bayesian Networks used in real Decision Support Systems.



Results

	MB with one spouse, three parents, and two children				MB with four spouses, one parent, and two children			
	50 Vars	200 Vars	1000 Vars	Average	50 Vars	200 Vars	1000 Vars	Average
IAMB	94.53	91.00	91.43	92.32	85.05	87.11	87.90	86.68
interIAMB	91.93	91.00	91.43	91.46	85.05	87.11	87.90	86.68
interIAMBnPC	93.67	94.43	88.77	92.29	87.71	88.01	73.69	83.14
IAMBnPC	94.43	91.60	91.67	92.57	90.48	85.63	85.70	87.27
GS	86.36	90.46	83.07	86.63	74.58	74.57	73.51	74.22
KS, k=0	95.93	96.17	96.15	96.08	74.72	73.39	73.06	73.72
KS, k=1	79.91	71.13	73.37	74.80	85.94	79.92	79.08	81.65
KS, k=2	86.11	87.35	86.94	86.80	85.88	82.24	81.40	83.17
PC	95.60	-	-	-	96.43	-	-	-

Table 2: Experiments on randomly generated Bayesian Networks



Outline

- Motivation
- Introduction to Bayesian Networks and Markov Blankets
- Markov Blanket Discovery algorithms
- IAMB algorithm and results
- HITON algorithm and results



HITON algorithm (Aliferis, *et al.*)

- Uses MB discovery technique for feature selection.
- Algorithm:
 - Identify the Markov Blanket of target T given the data D
 - Use wrapping to remove variables, which are unnecessary for predicting the target T , given algorithm A .
 - Return the minimal set (for predicting target T using algorithm A)



HITON algorithm

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

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

Return Vars

HITON-MB(Data D , Target T)

“returns the Markov Blanket of T ”

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

$PCPC$ = parents and children of the parents and children of T

$\text{CurrentMB} = PC \cup PCPC$

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

\forall potential spouse $X \in \text{CurrentMB}$ and $\forall Y \in PC$:

if $\neg \exists S \subseteq \{Y\} \cup V - \{T, X\}$ so that $\perp (T; X | S)$

then retain X in CurrentMB

else remove it

Return CurrentMB



HITON-PC(Data D, Target T)

“returns parents and children of T ”

$CurrentPC = \{\}$

Repeat

Find variable $V_i \notin CurrentPC$ that maximizes $association(V_i, T)$ 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 in phase I

Until no more variables are left to consider

Return $CurrentPC$

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 algorithm

- Aim: Good variable selection with the given classification algorithm, i.e., HITON employs a wrapper approach.
- First identify the MB(T), then remove any variables not required for classification given a classifier.



Results (Data-Sets)

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

A variety of biomedical tasks with different characteristics



Results

- Evaluation metric:
 - Area under the ROC curve.
- Results:
 - HITON consistently produces the smallest variable sets.
 - It exhibits best classification performance, and maximum variable reduction.

Results

1. Drug Discovery (Thrombin)					4. Gene Expression Diagnosis (Lung Cancer)				
	UAF*	RFE	HITON	ALL		UAF*	RFE*	HITON*	ALL*
SVM	96.12%	93.29%	93.23%	93.69%	SVM	99.32%	98.57%	97.83%	99.07%
KNN	87.25%	89.71%	92.23%	88.21%	NN	99.63%	98.70%	98.92%	N/A
NN	N/A	92.04%	92.65%	N/A	KNN	95.57%	91.49%	96.06%	97.59%
Average	91.69%	91.68%	92.7%	90.95%	Average	98.17%	96.25%	97.60%	98.33%
# of variables	34837	8709	32	139351	# of variables	330	19	16	12,600
2. Clinical Diagnosis (Arrythmia)					5. Mass-Spectrometry Diagnosis (Prostate Cancer)				
	UAF*	B/F*	HITON*	ALL*		UAF*	RFE*	HITON*	ALL*
DTI	73.94%	72.85%	71.87%	73.94%	SVM	98.50%	98.95%	99.10%	99.40%
KNN	63.22%	63.45%	65.30%	63.22%	NN	98.62%	98.78%	97.95%	99.27%
NN	58.29%	60.90%	60.38%	58.29%	KNN	77.52%	86.53%	91.36%	76.94%
Average	65.15%	65.73%	65.85%	65.15%	Average	91.55%	94.75%	96.14%	91.87%
# of variables	279	96	63	279	# of variables	706	87	16	779



Questions / Discussion



References

- [1] Aliferis, C., Tsamardinos, I., Statnikov A. (2003) HITON, A Novel Markov Blanket Algorithm for Optimal Variable Selection.
- [2] Bai X., *et al.* (2004) PCX: Markov Blanket Classification for Large Data Sets with Few Cases. *CMU-CALD*.
- [3] Koller D., Sahami M. (1996) Toward Optimal Feature Selection. *International Conference on Machine Learning*, pp. 284-292.
- [4] Tsamardinos, I., Aliferis, C., Statnikov A. (2003) Algorithms for Large Scale Markov Blanket Discovery. *The 16th International FLAIRS Conference*, St. Augustine, Florida, USA.
- [5] Yaramakala S. (2004) Fast Markov Blanket Discovery. MS – thesis, Iowa State University.
- [6] Yu, L., Liu, H. (2003) Feature Selection for High Dimensional Data: A Fast Correlation Based Filter Solution. ICML – 2003.