HITON: A Novel Markov Blanket Algorithm for Optimal Variable Selection

C.F. Aliferis M.D., Ph.D., I. Tsamardinos Ph.D., A. Statnikov M.S.

Department of Biomedical Informatics, Vanderbilt University

AMIA Fall Conference, November 2003

Feature Selection Problem Definition

- Informally: find smallest set of predictors with maximum classification performance

Importance of Feature Selection

- Essential component of construction of decision support models, and computer-assisted discovery.
  - In medical diagnosis, for example, elimination of redundant tests from consideration reduces risks to patients and lowers healthcare costs.
  - May Improve performance of classification algorithm
  - Classification algorithm may not scale up to the size of the full feature set either in sample or time
  - Allows us to better understand the domain

Importance of Feature Selection

- Problem of variable selection in biomedicine is more pressing than ever, due to the recent emergence of extremely large datasets, sometimes involving tens to hundreds of thousands of variables. Some examples:
  - gene-expression array studies,
  - proteomics,
  - computational biology,
  - text-categorization,
  - information retrieval,
  - mining of electronic medical records,
  - consumer profile analysis,
  - temporal modelling, etc.

Approaches

- “Wrapping”
  - conduct heuristic search in space of all possible variable subsets (e.g., greedy search, branch-and-bound, genetic algorithm search)
  - evaluate each subset by building a model (using the classifier of choice) and using cross-validation to estimate its future classification performance
  - return best subset

Approaches

- “Filtering”
  - Do not evaluate directly the classifier with multiple feature subsets; instead, typically, use properties of the distribution of variables to identify potentially good predictors (e.g., choose predictors that are most strongly associated with the target, or use predictors that are grouped together with the target by a clustering algorithm)

Wrapping or Filtering?

- Wrappers are classifier-specific and thus if one could do a thorough search they would return the optimal solution. Unfortunately the search space is so large that search is always incomplete.
- No single wrapper algorithm is best over all possible classification tasks! (Tsamardinos and Aliferis 2003)
- We can only prove that a *specific* filter (or wrapper) algorithm for a *specific* classifier (or class of classifiers), a specific loss function, and a *specific* class of distributions yields optimal or sub-optimal solutions; and unless we provide such proofs we are operating on faith and hope…

The Markov Blanket as Solution to the Variable Selection Problem

- Several researchers (Cooper et al., Koller et al., Cheng et al.) have suggested, intuitively, that the Markov Blanket of the target variable *T,* denoted as *MB(T),* is a key concept for solving the variable selection problem.
- Markov Blanket (of a variable *T*): the set of variables MB, such that conditioned on MB every other variable becomes independent of *T*.
- Thus, intuitively, knowledge of the values of the Markov Blanket variables should render all other variables superfluous for classifying *T*.

The Markov Blanket as Solution to the Variable Selection Problem

- Tsamardinos and Aliferis showed recently (AI and Statistics 2003) that:
  - in distributions the dependencies and independencies of which can be captured by a Bayesian Network (“Faithful” distributions)
  - the Markov Blanket is unique, and that
  - MB is precisely the solution to the feature selection problem as long as we use a sufficiently powerful classifier class, and quadratic loss functions

How hard is the Markov Blanket learning problem?

- This is a hard problem given the limitations of existing algorithms:
  - K2MB algorithm (Cooper et al): unsound, does not scale up
  - Koller and Sahami: unsound, does not scale up
  - Margaritis and Thrun: sound requires sample exponential to size of MB
  - Tsamardinos, Aliferis, Statnikov (previous work): sound requires sample exponential to size of MB
  - Cheng and Greiner: sound, does not scale up

How hard is it to learn the Markov Blanket?

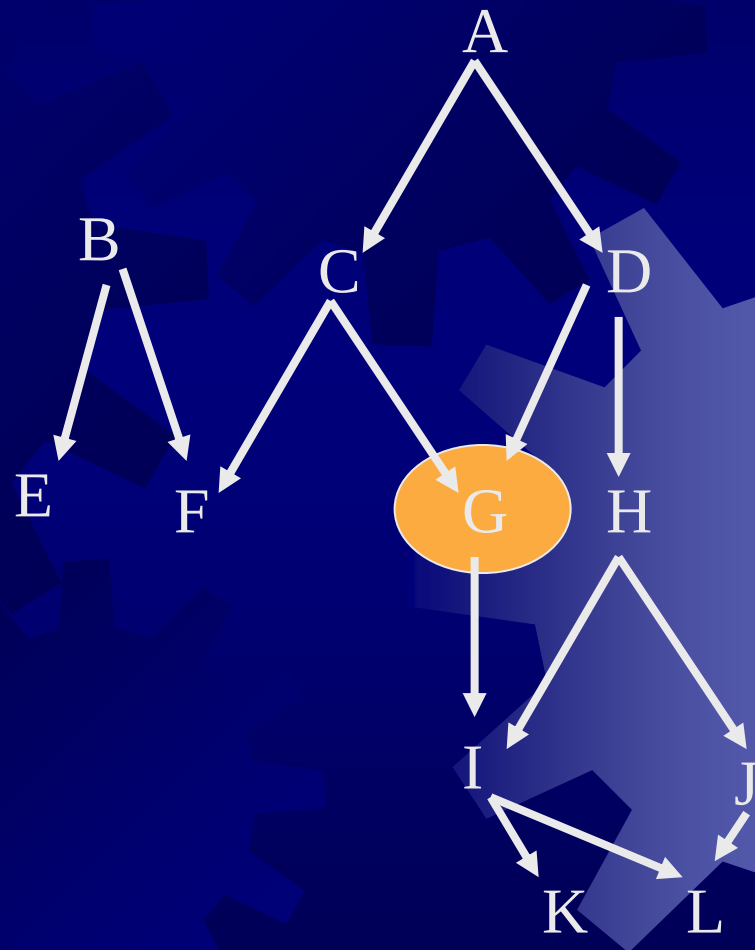
ALGORITHM	SOUND	SCALABLE	SAMPLE EXPONENTIAL TO |MB|	COMMENTS
Cheng and Greiner	YES	NO	NO	Post-processing on learning BN
Cooper et al.	NO	NO	NO	Uses full BN learning
Margaritis and Thrun	YES	YES	YES	Intended to facilitate BN learning
Koller and Sahami	NO	NO	NO	Most widely-cited MB induction algorithm
Tsamardinos and Aiferis	YES	YES	YES	Some use BN learning as sub-routine
Ideally →	YES	YES	NO	

This Research's Goals:

- To construct a new MB induction algorithm ("HITON") that is:
  - Sound
  - Highly scalable in practice
  - Does not require sample exponential to the size of MB
- To conduct a thorough initial evaluation of the algorithm's performance in terms of achieved classification and variable set shrinkage in several domains:
  - gene expression diagnosis,
  - mass spectrometry diagnosis,
  - molecular function prediction from structural properties,
  - clinical diagnosis, and
  - text categorization

Methods: algorithm

- **Step #1: Find the parents and children of *T;* call this set PC(*T*)**
- Step #2: Find the **PC(.)** set of each member of **PC(*T*)** using conditional independence tests**;** take the union of all these sets to be **PCunion**
- Step #3: Run a special test to filter out from **PCunion** the non-members of MB(*T*)
- Step #4: Apply heuristic search with a desired classifier/loss function and cross-validation to identify variables that can be dropped from MB(T) without loss of accuracy



Methods: algorithm properties

ALGORITHM	SOUND	SCALABLE	SAMPLE EXPONENTIAL TO |MB|
Cheng and Greiner	YES	NO	NO
Cooper et al.	NO	NO	NO
Margaritis and Thrun	YES	YES	YES
Koller and Sahami	NO	NO	NO
Tsamardinos and Aiferis	YES	YES	YES
HITON	**YES**	**YES**	**NO**

Methods: datasets

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.

Figure 2: Dataset Characteristics

Methods: evaluation metrics

- The classifiers' outputs were thresholded to derive the ROCs.
- AUC was computed using the trapezoidal rule and statistical comparisons among AUCs were performed using an unpaired Wilcoxon rank sum test.
- The size reduction was evaluated by fraction of variables in the resulting models.
- All metrics (variable reduction, AUC) were averaged over cross-validation splits.

Methods: Statistical choices and cross-validation

- Statistical choices. In all our experiments conditional tests of independence were implemented as G^2 tests with a significance level set to 5%, and degrees of freedom according to Spirtes et al (2000).
- Cross-validation. We employed a nested stratified cross-validation design throughout, in which the outer loop of cross-validation estimates the performance of the optimised classifiers while the inner loop is used to find the best parameter configuration/variable subset for each classifier.

Methods: Classifiers

- We applied several state-of-the-art classifiers:
  - Polynomial-kernel Support Vector Machines (SVM),
  - K-Nearest Neighbors (KNN),
  - Feed-forward Neural Networks (NNs),
  - Decision Trees (DTI), and
  - Simple Bayes Classifier (text categorization).

Methods: Classifiers

- Details of application of the classifiers:
  - For SVMs we used the LibSVM implementation that is based on Platt's SMO algorithm , with C chosen from the set: {10^{-8}, 10^{-7}, 10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 0.1, 1, 10, 100, 1000} and degree from the set: {1, 2, 3, 4}.
  - For KNN, we chose k from the range: [1,…, number of samples in the training set] using our own implementation of the algorithm.
  - For NNs we used the Matlab Neural Network Toolbox with 1 hidden layer, number of units chosen (heuristically) from the set {2, 3, 5, 8, 10, 30, 50}, variable-learning-rate back propagation, custom-coded early stopping with (limiting) performance goal=10^{-8} (i.e., an arbitrary value very close to zero), and number of epochs in the range [100,…,10000], and a fixed momentum of 0.001.
  - We used Quinlan's See5 commercial implementation of C4.5 Decision Tree Induction and our own implementation of the text-oriented Simple Bayes Classifier described in Mitchell (1997).

Methods: feature selection baselines

- Univariate Association Filtering (UAF) (for all tasks). UAF is a common and robust applied classical statistics procedure. In text categorization especially, extensive experiments have established its superior performance [25]. UA. We used Fisher Criterion Scoring for gene expression data [3], X^2 and Information Gain for text categorization, Kruskal-Wallis ANOVA for the continuous variables of Arrhythmia, and G^2, for the remaining datasets.
- Recursive Feature Elimination (RFE) (for bioinformatics- related tasks),
- Backward/Forward Wrapping (for the clinical diagnosis task).

Results

1. Drug Discovery (Thrombin)				
	UAF*	RFE	HITON	ALL
SVM	96.12%	93.29%	93.23%	93.69%
KNN	87.25%	89.71%	92.23%	88.21%
NN	*NA*	92.04%	92.65%	*NA*
Average	91.69%	91.68%	**92.7%**	90.95%
# of variables	34837	8709	**32**	139351

2. Clinical Diagnosis (Arrythmia)				
	UAF*	B/F*	HITON*	ALL*
DTI	73.94%	72.85%	71.87%	73.94%
KNN	63.22%	63.45%	65.30%	63.22%
NN	58.29%	60.90%	60.38%	58.29%
Average	65.15%	65.73%	**65.85%**	65.15%
# of variables	279	96	**63**	279

3. Text Categorization (OHSUMED)				
	IG	X^2	HITON	ALL*
SVM	82.43%	85.91%	82.85%	90.50%
SBCtc	84.18%	86.23%	85.10%	84.25%
KNN	75.55%	81.76%	80.25%	77.56%
NN	82.47%	85.27%	83.97%	*NA*
Average	81.16%	**84.79%**	83.04%	84.10%
# of variables	224	112	**34**	14373

4. Gene Expression Diagnosis (Lung Cancer)				
	UAF*	RFE*	HITON*	ALL*
SVM	99.32%	98.57%	97.83%	99.07%
NN	99.63%	98.70%	98.92%	N/A
KNN	95.57%	91.49%	96.06%	97.59%
Average	98.17%	96.25%	97.60%	**98.33%**
# of variables	330	19	**16**	12,600
5. Mass-Spectrometry Diagnosis (Prostate Cancer)				
	UAF*	RFE*	HITON*	ALL*
SVM	98.50%	98.95%	99.10%	99.40%
NN	98.62%	98.78%	97.95%	99.27%
KNN	77.52%	86.53%	91.36%	76.94%
Average	91.55%	94.75%	**96.14%**	91.87%
# of variables	706	87	**16**	779
Averages Over All Tasks				
	Av. Over BaselineAlgorithms		HITON	ALL
Av. Perf. over classifiers	86.1%		**87.1%**	86.1%
Av. variable #	4540		**32.3**	33,476
Av. reduction	x 8		x **1124**	x 1

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).

Summary results

- HITON consistently produces the smallest variable sets in each task/dataset; The reduction in variables ranges from 4.4 times (Arrhythmia) to 4,315 times (thrombin).

- Averaged over all classifiers and tasks/datasets, HITON exhibits best classification performance, and best variable reduction (~140 times smaller models on the average, than the baseline methods' average, and ~1100 times on the average smaller models than the average total number of variables).

Summary results

- In 3 out of 5 tasks HITON produces the best single classifier or a classifier that is statistically non-significantly different from the best (compared to 4 out of 5 *for all other baselines combined*);

- Averaged over all classifiers in each task/dataset, HITON produces the models with best classification performance in 4 our of 5 tasks;

- Compared to using all variables, HITON improves performance 2 times out of 5, while maintains performance another two times out of 5 and yields minimally worse performance in the remaining task (text categorization).

Summary results

- HITON can be run in a few hours for massive datasets using very inexpensive computer platforms. For example, it took 8 to 9 hours (depending on classifier) to run in the massive thrombin dataset (baselines: 4 to 4.7 hours) using a Intel Xeon 2.4 GHz computer with 2 GB of RAM.

Added benefits

- In addition to addressing the feature selection problem MB induction is very important for:
  - Discovery (since in faithful distributions, the MB is the set of direct causes,direct effects, and direct causes of the direct effects of the target T)
  - Learning the full causal network efficiently (by using MB induction in a divide-and-conquer manner and combining it with constraint-based or search-and-score methods)

Conclusions

- HITON is the first Markov Blanket – inducing algorithm that combines the following three properties:
  - (a) is sound;
  - (b) is highly-scalable to the number of variables;
  - (c) is sample-efficient relative to the size of the Markov Blanket.
- Our experimental evaluation suggests that it is applicable to a wide variety of biomedical data, in particular: structural molecular biology, clinical diagnosis, text-categorization, gene expression analysis, and proteomics.

Additional Information

- I. Tsamardinos, C.F. Aliferis, A. Statnikov. "Time and Sample Efficient Discovery of Markov Blankets and Direct Causal Relations" *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2003
- Tsamardinos I and C.F. Aliferis. Towards principled feature selection: relevancy, filters, and wrappers. *Proc. AI and Statistics*, 2003.

Conclusions

- Given that HITON has a well-specified set of assumptions for inducing the Markov Blanket:
  - Faithfulness
  - Reliable statistical tests of independence
- And additionally for the MB to solve the feature selection problem optimally:
  - Powerful classifiers
  - Loss function is accuracy or quadratic loss
- Especially with respect to the faithfulness assumption, HITON's robustness in our experiments implies that either biomedical data do not exhibit severe violations of this distributional assumption, or that such violations are mitigated by currently poorly-understood factors.