Machine Learning For Bioinformatics: Feature Selection Methods

C.F. Aliferis M.D., Ph.D.

9-17-2002

Discovery Systems Laboratory,
Department of Biomedical Informatics,
Vanderbilt University

Goals

- Quick reminder of topics covered last time
- What is Feature Selection for classification?
- Why feature selection is important?
- What is the filter and what is the wrapper approach to feature selection?
- Major Feature Selection Methods in Bioinformatics
- How do we approach the feature selection problem in DSL research?
- Example application in drug R&D

Topics Covered Previously

- What is *Machine Learning* (ML)? How is it different than *Statistics* and *Data Mining*?
- Example applications of ML in: drug development, bioinformatics, and clinical problem areas
- Difference between *supervised* and *unsupervised* ML methods
- Theoretical basis of supervised Inductive ML
- How can ML methods fail

Topics Covered Previously CNT'D

- Decision Tree Induction (Lab with See5)
- K-Nearest Neighbors
- Genetic Algorithms
- Artificial Neural Networks (Lab with NevProp)
- Clustering
- Causal probabilistic Network Induction (Lab with BN Power Constructor)

What is Feature Selection for classification?

- Given: a set of predictors (“features”) *V* and a target variable *T*
- Find: minimum set *F* that achieves maximum classification performance of *T*

Why feature selection is important?

- 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
- Cheaper to collect a reduced set of predictors
- Safer to collect a reduced set of predictors

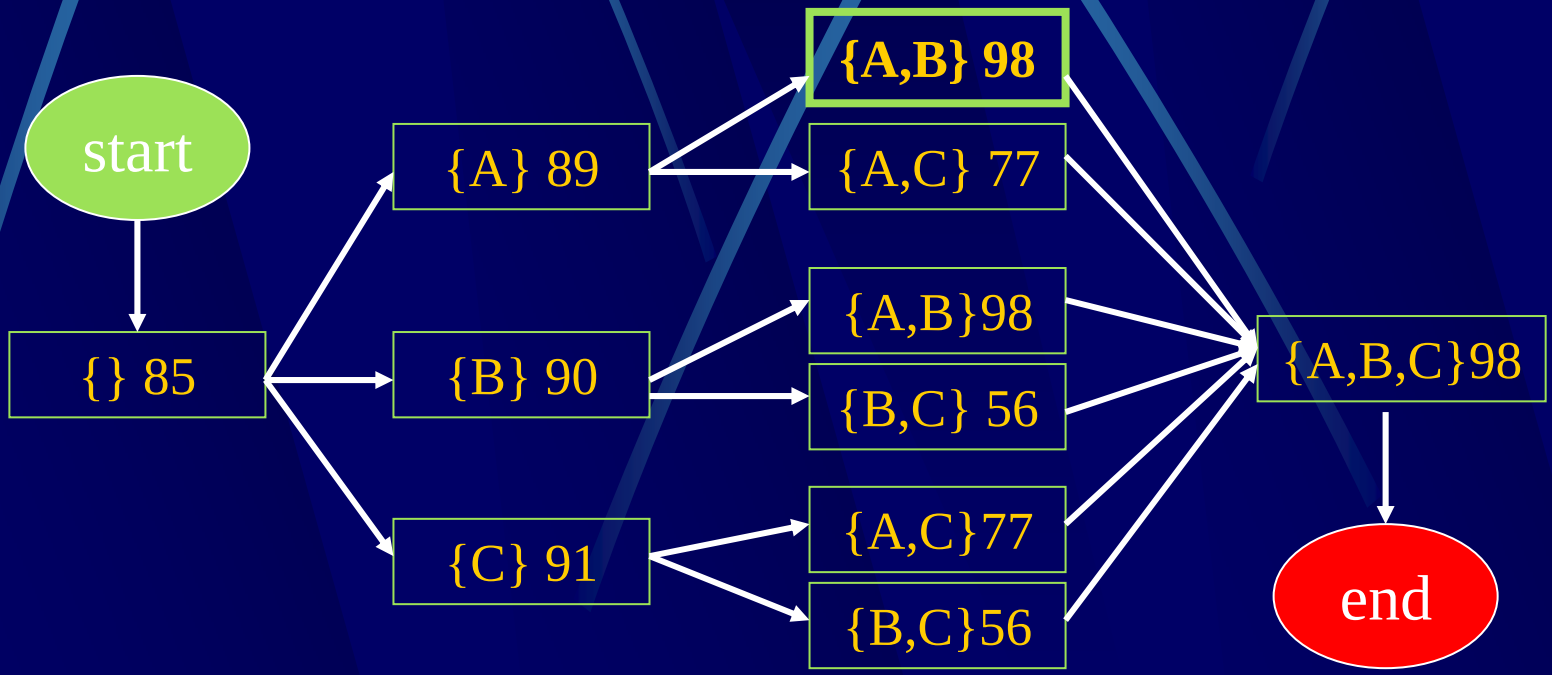
Filters vs Wrappers: Wrappers

Say we have predictors A, B, C and classifier *M*. We want to predict T given the smallest possible subset of {A,B,C}, while achieving maximal performance

FEATURE SET	CLASSIFIER	PERFORMANCE
{A,B,C}	*M*	**98%**
{A,B}	*M*	**98%**
{A,C}	*M*	77%
{B,C}	*M*	56%
{A}	*M*	89%
{B}	*M*	90%
{C}	*M*	91%
{.}	*M*	85%

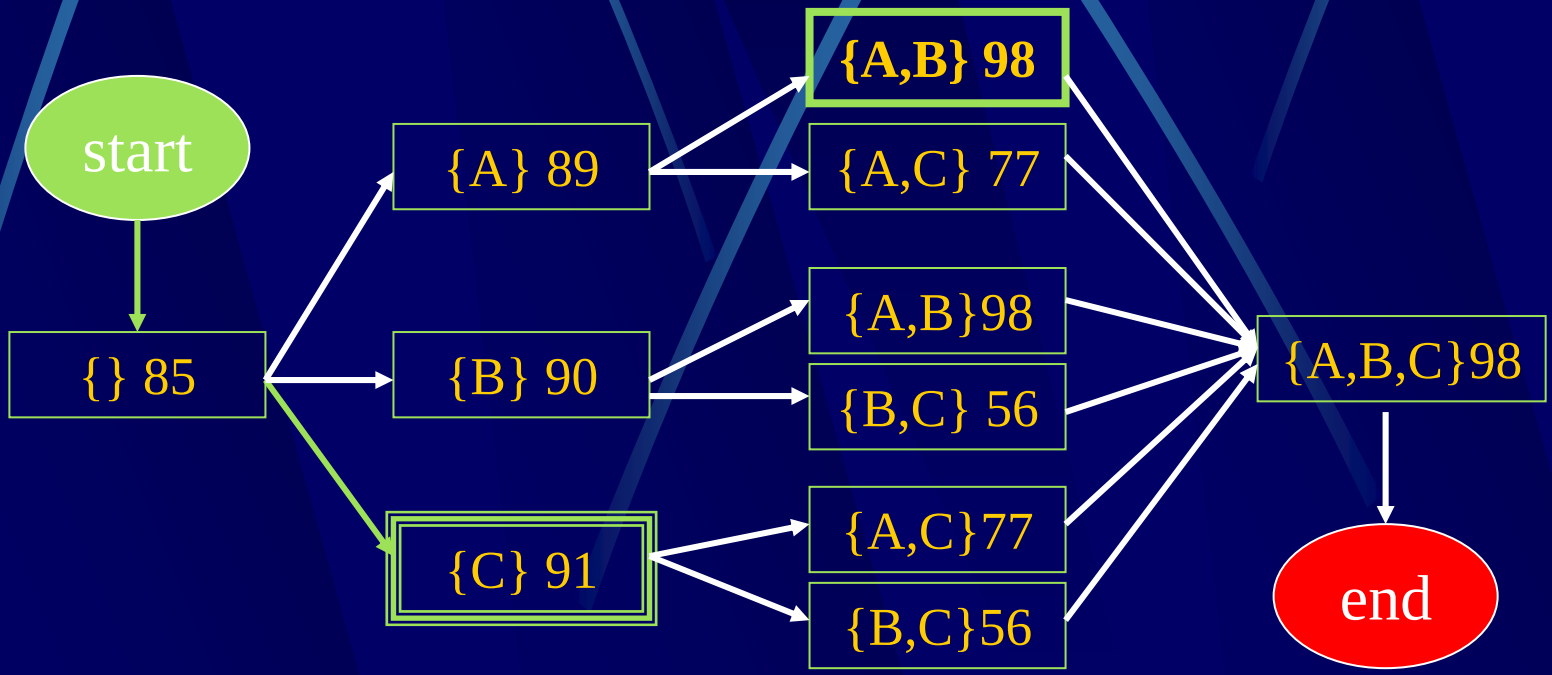
Filters vs Wrappers: Wrappers

The set of all subsets is the power set and its size is $2^{|V|}$.
Hence for large *V* we cannot do this procedure exhaustively; instead we rely on *heuristic search* of the space of all possible feature subsets.



Filters vs Wrappers: Wrappers

A common example of heuristic search is hill climbing: keep adding features one at a time until no further improvement can be achieved.



Filters vs Wrappers: Filters

In the filter approach we do not rely on running a particular classifier and searching in the space of feature subsets; instead we select features on the basis of statistical properties. A classic example is univariate associations:

FEATURE	ASSOCIATION WITH TARGET	
{A}	89%	Threshold gives suboptimal solution
{B}	90%	Threshold gives optimal solution
{C}	91%	Threshold gives suboptimal solution

Major Feature Selection Methods in Bioinformatics: Univariate Association Filtering

- Order all predictors according to strength of association with target
- Choose the first *k* predictors and feed them to the classifier
- Various measures of association may be used: X^2, G^2, Pearson r, Fisher Criterion Scoring, etc.

Major Feature Selection Methods in Bioinformatics: Recursive Feature Elimination

- Filter algorithm where feature selection is done as follows:

1. build linear Support Vector Machine classifiers using *V* features
2. compute weights of all features and choose the best *V/2*
3. repeat until 1 feature is left
4. choose the feature subset that gives the best performance
5. give best feature set to the classifier of choice.

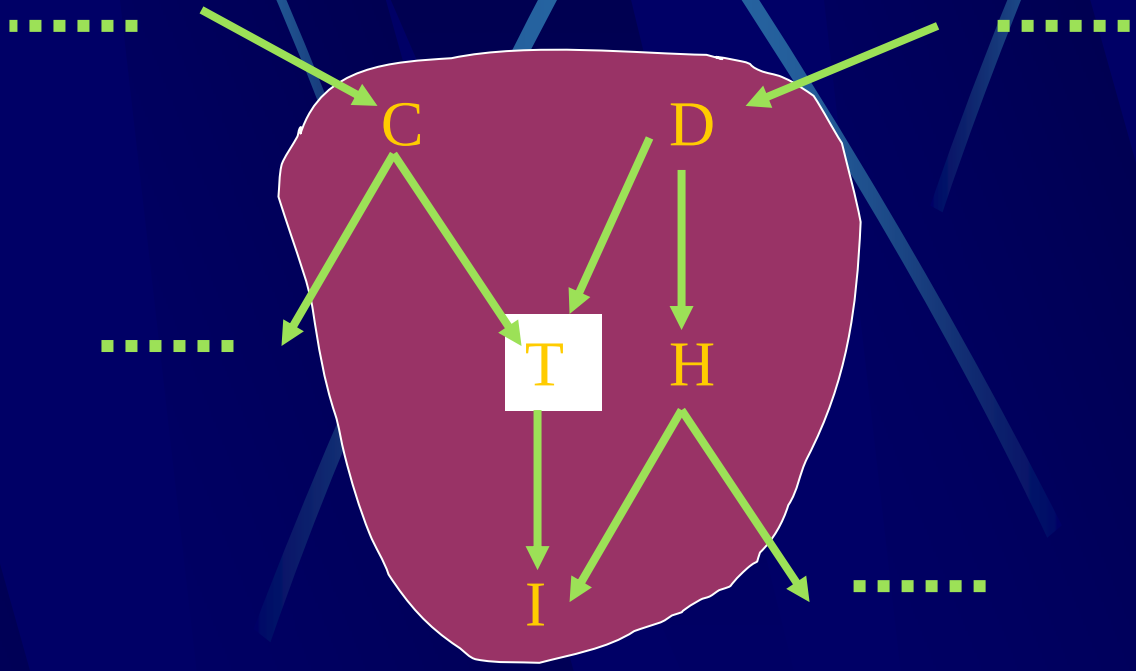
Major Feature Selection Methods in Bioinformatics: GA/KNN

- Wrapper approach whereby:

1. heuristic search=Genetic Algorithm, and
2. classifier=KNN

How do we approach the feature selection problem in DSL research?

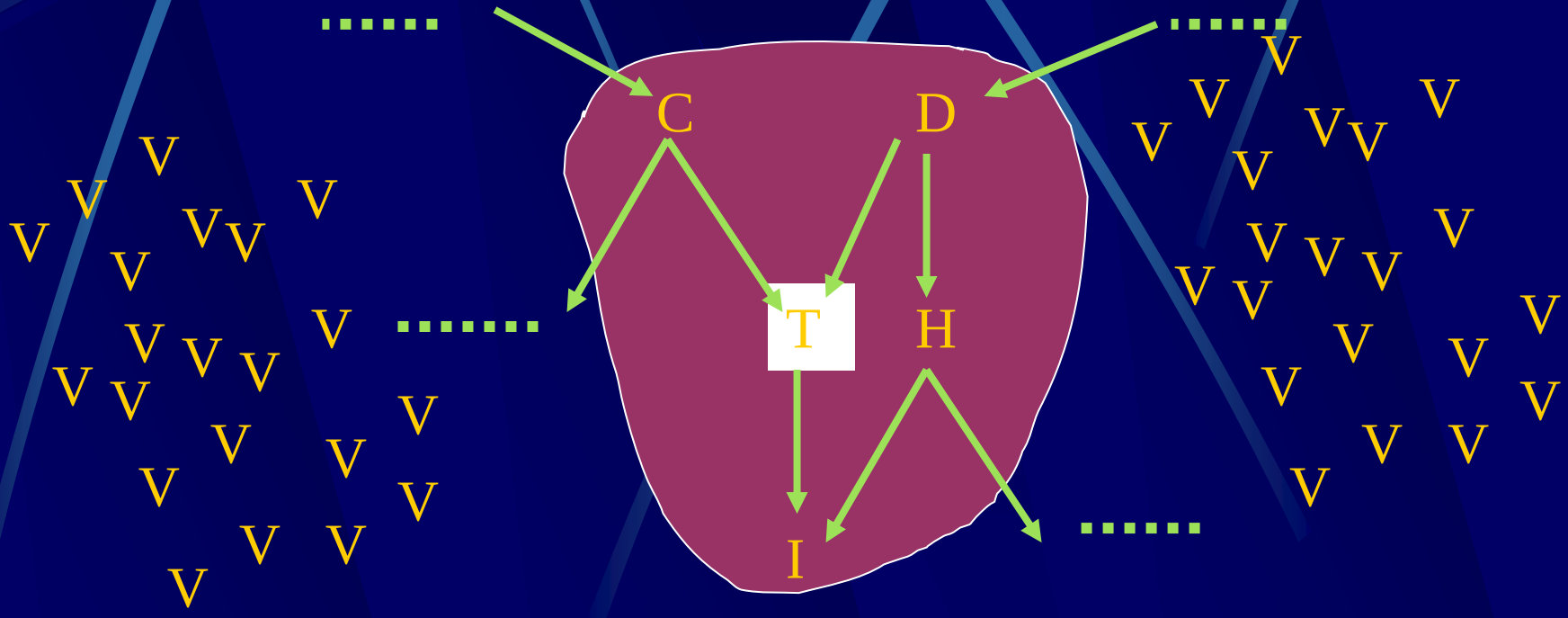
- (Reminder) Definition:
  - The Markov Blanket of some variable of interest **T** (“MB(T)”) is the set of the immediate causes, immediate effects, and immediate causes of the immediate effects of **T.**



Note: C causes T, T causes I ,etc.

A Crucial property of the Markov Blanket

- MB(T) is the minimal set of predictor variables needed for classification (diagnosis, prognosis, etc.) of the target variable **T**



So the Feature Selection Problem Statement Becomes:

- Goal:
  - Given: Data (observations of **T** and a set of variables V)
  - Find: MB(T)

Traditional MB Induction

- Previously MB(T) could be discovered using a full-network induction algorithm, or various heuristic procedures
- The state-of-the-art (full-network) algorithms try to learn the whole network and are not tractable for large networks

New Scalable Algorithms for Learning MB

Characteristics of newly-developed algorithms :

- Sound given broad and well-defined assumptions
- Scale up to hundreds of thousands of variables
- Quality of output insensitive to errors in learning about the rest of the variables
- Computational performance insensitive to structure beyond the target **T**
- Behave well when confounders are not observed

Example application in drug R&D

- Task: given 139,351 molecular structural properties classify molecules according to whether they bind to thrombin (and thus are good candidates as anti-clotting agents) [KDD Cup 2001, DuPont Pharmaceuticals]

Thrombin Task: Data Splits

DATA SET	SIZE (% OF split)	ACTIVE	INACTIVE	% OF ACTIVE IN SPLIT
TOTAL	2543 (-)	192	2351	7.6
TRAIN	2000 (78.7)	151	1849	7.6
TRAIN-TRAIN	1300 (65)	90	1210	6.9
VALIDATION	700 (35)	61	639	8.7
TEST	543 (21.3)	41	502	7.6

Thrombin Task: Performance

	IAMB (MI, Th=0.0143)	IAMB Chunked (MI, Th=0.0143, 14 chunks)	MMMB (MI, Th=0.01, |Cond|=5)	UAF	RFE	All
LSVM	88.3000%	93.4800%	94.8000%	94.7300%	93.2878%	93.4300%
PSVM	86.2600%	93.7800%	93.4400%	94.4600%	92.4716%	93.6900%
SBC	92.7500%	93.2500%	94.7700%	94.0500%	85.2128%	80.3300%
KNN	94.0600%	91.2100%	93.7900%	94.7800%	89.7095%	88.2100%
NN	94.1500%	93.3000%	93.5900%	88.8900%	92.0416%	N/A
Averages	91.1040%	93.0040%	94.0780%	93.3820%	90.5447%	88.9150%
Number of features	8	9	27	200	8709	139351

Thrombin Task: Parallelization

ALGORITHM	TOTAL TIME (HRS)	NOTES
LAMB	72	1 CPU, data in sparse array, 128MB RAM, 600 MHz PIII (100% load)
Chunked LAMB	6.69	1 CPU, data in dense array, 256MB RAM, 600 MHz PIII (100% load)
Fine-Grain Parallel LAMB	0.5	14 CPUs, data in dense arrays, 256MB RAM, 600 MHz PIII (100% load)
Fine-Grain Parallel LAMB distributed data	0.4	14 CPUs, data in dense arrays, 256MB RAM, 600 MHz PIII (100% load)
Chunked Parallel LAMB	0.53	14 CPUs, data in dense arrays, 256MB RAM, 600 MHz PIII (100% load)
Chunked Parallel LAMB distributed data	0.53	14 CPUs, data in dense arrays, 256MB RAM, 600 MHz PIII (100% load)

Filters vs Wrappers: Which Is Best?

- None over all possible classification tasks!
- We can only prove that a *specific* filter (or wrapper) algorithm for a *specific* classifier (or class of classifiers), and a *specific* class of distributions yields optimal or sub-optimal solutions. Unless we provide such proofs we are operating on faith and hope…

What is the biological significance of consistently selected features?

- In MB-based feature selection and CPN-faithful distributions: causal neighborhood of target (i.e., direct causes, direct effects, direct causes of the direct effects of target).
- In other methods: ???

References

- Feature Selection with RFE: **Gene Selection for Cancer Classification using Support Vector Machines (2000)** Isabelle Guyon, Jason Weston, Stephen Barnhill, Vladimir Vapnik; Machine Learning
- Feature Selection with GA/KNN: Li, Pedersen, Darden, and Weinberg (2001a). Computational analysis of leukemia microarray expression data using the GA/KNN method, Critical Assessment of Microarray Data Analysis 2001 (CAMDA'01)
- Feature Selection with Fisher Criterion: **Support Vector Machine Classification and Validation of Cancer Tissue Samples Using Microarray Expression Data (2001)** Terrence S. Furey, Nello Cristianini, Nigel Duffy, David W. Bednarski, Michèl Schummer, David Haussler Bioinformatics