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.

<table>
<thead>
<tr>
<th>FEATURE SET</th>
<th>CLASSIFIER</th>
<th>PERFORMANCE</th>
</tr>
</thead>
<tbody>
<tr>
<td>{A,B,C}</td>
<td>$M$</td>
<td>98%</td>
</tr>
<tr>
<td>{A,B}</td>
<td>$M$</td>
<td>98%</td>
</tr>
<tr>
<td>{A,C}</td>
<td>$M$</td>
<td>77%</td>
</tr>
<tr>
<td>{B,C}</td>
<td>$M$</td>
<td>56%</td>
</tr>
<tr>
<td>{A}</td>
<td>$M$</td>
<td>89%</td>
</tr>
<tr>
<td>{B}</td>
<td>$M$</td>
<td>90%</td>
</tr>
<tr>
<td>{C}</td>
<td>$M$</td>
<td>91%</td>
</tr>
<tr>
<td>{}</td>
<td>$M$</td>
<td>85%</td>
</tr>
</tbody>
</table>
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.
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:

<table>
<thead>
<tr>
<th>FEATURE</th>
<th>ASSOCIATION WITH TARGET</th>
</tr>
</thead>
<tbody>
<tr>
<td>{A}</td>
<td>89%</td>
</tr>
<tr>
<td>{B}</td>
<td>90%</td>
</tr>
<tr>
<td>{C}</td>
<td>91%</td>
</tr>
</tbody>
</table>

Threshold gives suboptimal solution

Threshold gives optimal solution

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

<table>
<thead>
<tr>
<th>DATA SET</th>
<th>SIZE (% OF split)</th>
<th>ACTIVE</th>
<th>INACTIVE</th>
<th>% OF ACTIVE IN SPLIT</th>
</tr>
</thead>
<tbody>
<tr>
<td>TOTAL</td>
<td>2543 (-)</td>
<td>192</td>
<td>2351</td>
<td>7.6</td>
</tr>
<tr>
<td>TRAIN</td>
<td>2000 (78.7)</td>
<td>151</td>
<td>1849</td>
<td>7.6</td>
</tr>
<tr>
<td>TRAIN-TRAIN</td>
<td>1300 (65)</td>
<td>90</td>
<td>1210</td>
<td>6.9</td>
</tr>
<tr>
<td>VALIDATION</td>
<td>700 (35)</td>
<td>61</td>
<td>639</td>
<td>8.7</td>
</tr>
<tr>
<td>TEST</td>
<td>543 (21.3)</td>
<td>41</td>
<td>502</td>
<td>7.6</td>
</tr>
</tbody>
</table>
## Thrombin Task: Performance

<table>
<thead>
<tr>
<th></th>
<th>IAMB (MI, Th=0.0143)</th>
<th>IAMB Chunked (MI, Th=0.0143, 14 chunks)</th>
<th>MMB (MI, Th=0.01, Cond=5)</th>
<th>UAF</th>
<th>RFE</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSVM</td>
<td>88.3000%</td>
<td>93.4800%</td>
<td>94.6000%</td>
<td>94.7300%</td>
<td>93.2873%</td>
<td>93.4300%</td>
</tr>
<tr>
<td>PSVM</td>
<td>86.2500%</td>
<td>93.7800%</td>
<td>93.4400%</td>
<td>94.4600%</td>
<td>92.4712%</td>
<td>93.6900%</td>
</tr>
<tr>
<td>SBC</td>
<td>92.7500%</td>
<td>93.2500%</td>
<td>94.7700%</td>
<td>94.0500%</td>
<td>85.2129%</td>
<td>90.3300%</td>
</tr>
<tr>
<td>KNN</td>
<td>94.0600%</td>
<td>91.2100%</td>
<td>93.7900%</td>
<td>94.7800%</td>
<td>89.7095%</td>
<td>88.2100%</td>
</tr>
<tr>
<td>NN</td>
<td>94.1500%</td>
<td>93.3000%</td>
<td>93.5900%</td>
<td>88.3900%</td>
<td>92.0415%</td>
<td>N/A</td>
</tr>
<tr>
<td><strong>Averages</strong></td>
<td><strong>91.1340%</strong></td>
<td><strong>93.0040%</strong></td>
<td><strong>94.0760%</strong></td>
<td><strong>93.3620%</strong></td>
<td><strong>90.5447%</strong></td>
<td><strong>68.9150%</strong></td>
</tr>
<tr>
<td><strong>Number of features</strong></td>
<td>8</td>
<td>9</td>
<td>27</td>
<td>200</td>
<td>8709</td>
<td>139351</td>
</tr>
</tbody>
</table>
# Thrombin Task: Parallelization

<table>
<thead>
<tr>
<th>ALGORITHM</th>
<th>TOTAL TIME (HRS)</th>
<th>NOTES</th>
</tr>
</thead>
</table>
| 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 array, 256MB RAM, 600 MHz  
                             |                  | PIII (100% load)                                                      |
| Fine-Grain Parallel LAMB       | 0.4              | 14 CPUs, data in dense array, 256MB RAM, 600 MHz  
                             | distributed data   | PIII (100% load)                                                      |
| Chunked Parallel LAMB          | 0.53             | 14 CPUs, data in dense array, 256MB RAM, 600 MHz  
                             |                  | PIII (100% load)                                                      |
| Chunked Parallel LAMB          | 0.53             | 14 CPUs, data in dense array, 256MB RAM, 600 MHz  
                             | distributed data   | 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 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