Towards robust feature selection for high-dimensional, small sample settings

Yvan Saeys

Bioinformatics and Evolutionary Genomics, Ghent University, Belgium
yvan.saeys@psb.ugent.be

Marseille, January 14th, 2010
Background: biomarker discovery

- Common task in computational biology
- Find the entities that best explain phenotypic differences
- Challenges:
  - Many possible biomarkers (high dimensionality)
  - Only very few biomarkers are important for the specific phenotypic difference
  - Very few samples
- Examples:
  - Microarray data
  - Mass spectrometry data
  - SNP data
Dimensionality reduction techniques

- Feature selection techniques
  - Subset selection
  - Feature ranking
  - Feature weighting

- Feature transformation techniques
  - Projection
    - PCA
    - LDA
  - Compression
    - Fourier transform
    - Wavelet transform
Dimensionality reduction techniques

Feature selection techniques
- Subset selection
- Feature ranking
- Feature weighting

Feature transformation techniques
- Projection
  - PCA
  - LDA
- Compression
  - Fourier transform
  - Wavelet transform

Preserve the original semantics!
Casting the problem as a feature selection task

- Feature selection is a way to avoid the curse of dimensionality
- Improve model performance
  - Classification: improve classification performance (maximize accuracy, AUC)
  - Clustering: improve cluster detection (AIC, BIC, sum of squares, various indices)
  - Regression: improve fit (sum of squares error)
- Faster and more cost-effective models
- Improve generalization performance (avoiding overfitting)
- Gain deeper insight into the processes that generated the data (esp. important in Bioinformatics)
The need for robust marker selection algorithms
The need for robust marker selection algorithms
The need for robust marker selection algorithms

Ranked gene list:
• gene A
• gene B
• gene C
• gene D
• gene E
• ...

Yvan Saeys (UGent)
The need for robust marker selection algorithms

Ranked gene list:
- gene A
- gene B
- gene C
- gene D
- gene E
- ...

Ranked gene list:
- gene X
- gene A
- gene W
- gene Y
- gene C
- ...

Yvan Saeys (UGent)
The need for robust marker selection algorithms

Motivation

- Highly variable marker ranking algorithms decrease the confidence of a domain expert
  - Need to quantify the stability of a ranking algorithm
  - Use this as an additional criterion next to the predictive power
- More robust rankings yield a higher chance of representing biologically relevant markers
- Focus on quantifying/increasing marker stability within one data source
Formalizing feature selection robustness

Definition

Consider a dataset \( \mathcal{D} = \{ \mathbf{x}_1, \ldots, \mathbf{x}_M \} \), \( \mathbf{x}_i = (x_{i1}^1, \ldots, x_{iN}^N) \) with \( M \) instances and \( N \) features. A feature selection algorithm can then be defined as a mapping \( \mathcal{F} : \mathcal{D} \rightarrow \mathbf{f} \) from \( \mathcal{D} \) to an \( N \)-dimensional vector \( \mathbf{f} = (f_1, \ldots, f_N) \),

1. weighting: \( f_i = w_i \) denotes the weight of feature \( i \)
2. ranking: \( f_i \in \{1, 2, \ldots, N\} \) denotes the rank of feature \( i \)
3. subset selection: \( f_i = 0/1 \) denotes the exclusion/inclusion of feature \( i \) in the selected subset
Formalizing feature selection robustness

Research questions:

1. How stable are current feature selection techniques for high dimensional, small sample settings?
   - Analyze sensitivity of robustness to signature size, model parameters.

2. Can we increase the robustness of feature selection in this setting?

Definition

A feature selection algorithm is *stable* if small variations in the input [training data] result in small variations in the output [selected features]:

\[ \mathcal{F} \text{ is stable iff for } D \approx D', \text{ it follows that } S(f, f') < \epsilon \]

Methodological requirements:

1. Framework to generate small changes in training data
2. Similarity measures for feature weightings/rankings/subsets
Generating training set variations

A subsampling approach: Draw \( k \) subsamples of size \( \lceil xM \rceil \) (\( 0 < x < 1 \)) randomly without replacement from \( D \), where the parameters \( k \) and \( x \) can be varied.

In our experiments:

- \( k = 500 \)
- \( x = 0.9 \)

Algorithm

1. Generate \( k \) subsamples of size \( xM \), \( \{D_1, \ldots, D_k\} \)
2. Perform the basic feature selector \( F \) on each of these \( k \) subsamples
   \[
   \forall k : F(D_k) = f_k
   \]
3. Perform all \( \frac{k(k-1)}{2} \) pairwise comparisons, and average over them
   \[
   \text{Stab}(F) = \frac{2 \sum_{i=1}^{k} \sum_{j=i+1}^{k} S(f_i, f_j)}{k(k-1)}
   \]

where \( S(\cdot, \cdot) \) denotes an appropriate similarity function between weightings/rankings/subsets
Similarity measures for feature selection outputs

1. Weighting (Pearson CC):

   \[ S(f_i, f_j) = \frac{\sum_l (f_i^l - \mu_{f_i})(f_j^l - \mu_{f_j})}{\sqrt{\sum_l (f_i^l - \mu_{f_i})^2 \sum_l (f_j^l - \mu_{f_j})^2}} \]

2. Ranking (Spearman rank CC):

   \[ S(f_i, f_j) = 1 - 6 \sum_l \frac{(f_i^l - f_j^l)^2}{N(N^2 - 1)} \]

3. Subset selection (Jaccard index):

   \[ S(f_i, f_j) = \frac{|f_i \cap f_j|}{|f_i \cup f_j|} = \frac{\sum_l I(f_i^l = f_j^l = 1))}{\sum_l I(f_i^l + f_j^l > 0)} \]
Kuncheva’s index for comparing feature subsets

**Definition**
Let $A$ and $B$ be subsets of features, both of the same cardinality $s$. Let $r = |A \cap B|

Requirements for a desirable stability index for feature subsets:

1. **Monotonicity**: for a fixed subset size $s$, and number of features $N$, the larger the intersection between the subsets, the higher the value of the consistency index.

2. **Limits**: index should be bound by constants that do not depend on $N$ or $s$. Maximum should be attained when the subsets are identical: $r = s$

3. **Correction for chance**: index should have a constant value for independently drawn subsets of the same cardinality $s$. 
Kuncheva’s index for comparing feature subsets

General form of the index:

\[
\frac{\text{Observed } r - \text{Expected } r}{\text{Maximum } r - \text{Expected } r}
\]

For randomly drawn \( A \) and \( B \), the number of objects from \( A \) selected also in \( B \) is a random variable \( Y \) with hypergeometric distribution with probability mass function

\[
P(Y = r) = \binom{s}{r} \binom{N-s}{s-r} / \binom{N}{s}
\]

The expected value of \( Y \) for given \( s \) and \( N \) is \( \frac{s^2}{N} \). Thus define

\[
K_I(A, B) = \frac{r - \frac{s^2}{N}}{s - \frac{s^2}{N}} = \frac{rN - s^2}{s(N - s)}
\]

\( K_I \) is bound by \(-1 \leq K_I \leq 1 \) [Kuncheva (2007)]
Improving feature selection robustness

Methodology based on ensemble methods for classification. Can we transfer this to feature selection?

- Previous work
  - Use feature selection to construct an ensemble
  - Works of Cherkauer, Opitz, Tsymbal and Cunningham
  - Feature selection $\rightarrow$ ensemble

- This work
  - Use ensemble methods to perform feature selection
  - Feature selection $\leftarrow$ ensemble

Research questions:

- Can we improve feature selection robustness/stability using ensembles of feature selectors?
- Statistical, computational and representational aspects of ensemble learning transferable to feature selection?
- How does it affect classification performance?
Components of ensemble feature selection

Training set
Components of ensemble feature selection

- Training set
  - Feature selection algorithm 1
    - Ranked list 1
Components of ensemble feature selection

Training set

- Feature selection algorithm 1
  - Ranked list 1
- Feature selection algorithm 2
  - Ranked list 2
Components of ensemble feature selection

Training set

- Feature selection algorithm 1
  - Ranked list 1
- Feature selection algorithm 2
  - Ranked list 2
- Feature selection algorithm t
  - Ranked list T
Components of ensemble feature selection

Training set

- Feature selection algorithm 1
- Feature selection algorithm 2
- ... (Feature selection algorithm t)

- Ranked list 1
- Ranked list 2
- ... (Ranked list T)

Aggregation operator

Consensus

Ranked list C
Components of ensemble feature selection

- Variation in the feature selectors
  - Choosing different feature selection techniques
  - Dataset perturbation
    - *Instance level perturbation*
    - *Feature level perturbation*
  - *Stochasticity in the feature selector*
  - Bayesian model averaging
  - Combinations of these techniques
Components of ensemble feature selection

- Variation in the feature selectors
  - Choosing different feature selection techniques
  - Dataset perturbation
    - *Instance level perturbation*
    - *Feature level perturbation*
  - *Stochasticity in the feature selector*
  - Bayesian model averaging
  - Combinations of these techniques

- Aggregation of the results into a single output
  - *Rank aggregation*
  - *Weighted rank aggregations*
  - Score aggregation
  - *Counting most frequently selected features*
Overview: 2 case studies

1. Bagging based ensemble feature selection
   - Microarray data sets
   - Feature ranking approach
   - Rank aggregation method

2. Ensemble feature selection using model stochasticity
   - Mass spectrometry data sets
   - Feature selection approach
   - Subset aggregation approach
Case study 1: Bagging based ensemble feature selection

- Generate feature selection diversity by instance perturbation
  - Bootstrapping
  - Generate $t$ datasets by sampling the training set with replacement
  - For each dataset, apply a feature selection algorithm (e.g. a ranker)

$$EFS = \{F_1, F_2, \ldots, F_t\}$$

- Each feature selector $F_k$ results in a ranking $f_i = (f_i^1, \ldots, f_i^N)$, where $f_i^j$ denotes the rank of feature $j$ in bootstrap $i$. 
Aggregation methods

- Rank aggregation

\[
f = (\sum_{i=1}^{t} w_1 f_i^1, \ldots, \sum_{i=1}^{t} w_N f_i^N)
\]

- Complete linear aggregation (CLA)

\[w_i = 1\]

- Complete weighted aggregation (CWA)

\[w_i = \text{OO-AUC}_i\]
Overview methodology

Full data set (100% of the samples)

SUBSAMPLING

90 %

Marker selection algorithm

Ranked list 1

90 %

Marker selection algorithm

Ranked list 2

90 %

Marker selection algorithm

Ranked list K

…
Overview methodology

90 %

Marker selection algorithm

Ranked list 1
Overview methodology

BOOTSTRAPPING

90 %

Marker selection algorithm

Ranked list A

Marker selection algorithm

Ranked list B

Marker selection algorithm

Ranked list T

Bootstrap 1

Bootstrap 2

Bootstrap T

Yvan Saeys (UGent)
Towards robust feature selection
Marseille 2010 18 / 36
Overview methodology

Consensus marker selection algorithm

Consensus ranked list

90 %
Experiments

Microarray datasets

<table>
<thead>
<tr>
<th>Name</th>
<th># Class 1</th>
<th># Class 2</th>
<th>Size</th>
<th># Features</th>
<th>SDR</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Colon</td>
<td>40</td>
<td>22</td>
<td>62</td>
<td>2000</td>
<td>0.031</td>
<td>Alon et al. (1999)</td>
</tr>
<tr>
<td>Leukemia</td>
<td>47</td>
<td>25</td>
<td>72</td>
<td>7129</td>
<td>0.010</td>
<td>Golub et al. (1999)</td>
</tr>
<tr>
<td>Lymphoma</td>
<td>22</td>
<td>23</td>
<td>45</td>
<td>4026</td>
<td>0.011</td>
<td>Alizadeh et al. (2000)</td>
</tr>
<tr>
<td>Prostate</td>
<td>52</td>
<td>55</td>
<td>107</td>
<td>6033</td>
<td>0.017</td>
<td>Singh et al. (2002)</td>
</tr>
</tbody>
</table>

Baseline classifier/feature selection algorithm

- Linear SVM
- SVM Recursive Feature Elimination (RFE, Guyon et al. (2002))
  1. Train linear SVM on full feature set
  2. Rank features based on $|w|$
  3. Eliminate 50% worst features
  4. Retrain SVM on remaining features
  5. Go to step 2
Results: stability distributions
Results: stability

Colon

Leukemia

Lymphoma

Prostate

Towards robust feature selection
Results: classification performance

Colon

Leukemia

Lymphoma

Prostate

Yvan Saey (UGent)
Bagging based EFS: first conclusions

- Ensemble feature selection (EFS) increases model performance:
  - More stable biomarker selection
  - Increased predictive performance
- EFS is easy to parallelize
- As signature sizes get smaller, EFS progressively improves upon the baseline
- Robust, small signatures are interesting candidates for prognostic tests
- Linear aggregation method is preferred
Sensitivity analysis: number of bootstraps

Effect on stability

- Colon
- Leukemia
- Lymphoma
- Prostate

Yvan Saeys (UGent) Towards robust feature selection
Sensitivity analysis: number of bootstraps

Effect on classification performance

Colon

Leukemia

Lymphoma

Prostate

AUC

Percentage of selected features

AUC

Percentage of selected features

AUC

Percentage of selected features

AUC

Percentage of selected features

Yvan Saeys (UGent)
Towards robust feature selection
Marseille 2010
Sensitivity analysis: RFE elimination percentage

Effect on stability

Colon

Kuncheva Index

Percentage of selected features

Leukemia

Kuncheva Index

Percentage of selected features

Lymphoma

Kuncheva Index

Percentage of selected features

Prostate

Kuncheva Index

Percentage of selected features
Sensitivity analysis: RFE elimination percentage

Effect on classification performance

Colon

Leukemia

Lymphoma

Prostate

Yvan Saey (UGent)  Towards robust feature selection  Marseille 2010  27 / 36
Bagging based EFS: final conclusions

- Ensemble feature selection (EFS) increases model performance:
  - More stable biomarker selection
  - Increased predictive performance
- Number of bootstraps only effects stability
- RFE elimination percentage does not affect EFS
- RFE elimination percentage has a strong impact on baseline:
  - Single run SVM performs best in terms of stability
  - Smaller impact on classification performance
Case study 2: Ensemble FS using model stochasticity

- Traditional approach:
  - Run a stochastic FS method many times (e.g. MCMC, Genetic Algorithm, stochastic iterative sampling)
  - Compare all feature subsets found
  - Make a final selection
    - Intersection of the results
    - Most frequently selected features
Case study 2: Ensemble FS using model stochasticity

- **Traditional approach:**
  - Run a stochastic FS method many times (e.g. MCMC, Genetic Algorithm, stochastic iterative sampling)
  - Compare all feature subsets found
  - Make a final selection
    - Intersection of the results
    - Most frequently selected features

- **Computationally more efficient approach:**
  - Don’t use only the single best results of the sampling procedure
  - Average over the whole distribution
Estimation of distribution algorithms (EDA)

- Instead of working on one solution, work on a set of solutions (distribution)
- Use stochastic iterative sampling, combined with probabilistic graphical models to model good solutions

1. Generate initial solution set $S_0$
2. Select a number of samples
3. Estimate probability distribution
4. Generate new samples by sampling the estimated distribution
5. Create new solution set

Termination criteria met?
- No
- Yes
End

Yvan Saeys (UGent)
Estimating the probability distribution

Graphical model

**EDA**

**UMDA**

**BMDA**

**BOA, EBNA**

<table>
<thead>
<tr>
<th>$X_i$</th>
<th>$p(x_i^j)$, $j=0,1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_1$</td>
<td>$p(x_1^j)$</td>
</tr>
<tr>
<td>$X_2$</td>
<td>$p(x_2^j)$</td>
</tr>
<tr>
<td>$X_3$</td>
<td>$p(x_3^j)$</td>
</tr>
<tr>
<td>$X_4$</td>
<td>$p(x_4^j)$</td>
</tr>
<tr>
<td>$X_5$</td>
<td>$p(x_5^j)$</td>
</tr>
<tr>
<td>$X_6$</td>
<td>$p(x_6^j)$</td>
</tr>
<tr>
<td>$X_7$</td>
<td>$p(x_7^j)$</td>
</tr>
<tr>
<td>$X_8$</td>
<td>$p(x_8^j)$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$X_i$</th>
<th>$p(x_i^j)$, $j,k=0,1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_1$</td>
<td>$p(x_1^j)$</td>
</tr>
<tr>
<td>$X_2$</td>
<td>$p(x_2^j)$</td>
</tr>
<tr>
<td>$X_3$</td>
<td>$p(x_3^j)$</td>
</tr>
<tr>
<td>$X_4$</td>
<td>$p(x_4^j</td>
</tr>
<tr>
<td>$X_5$</td>
<td>$p(x_5^j</td>
</tr>
<tr>
<td>$X_6$</td>
<td>$p(x_6^j</td>
</tr>
<tr>
<td>$X_7$</td>
<td>$p(x_7^j</td>
</tr>
<tr>
<td>$X_8$</td>
<td>$p(x_8^j</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$X_i$</th>
<th>$p(x_i^j)$, $j,k,l=0,1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_1$</td>
<td>$p(x_1^j)$</td>
</tr>
<tr>
<td>$X_2$</td>
<td>$p(x_2^j)$</td>
</tr>
<tr>
<td>$X_3$</td>
<td>$p(x_3^j</td>
</tr>
<tr>
<td>$X_4$</td>
<td>$p(x_4^j</td>
</tr>
<tr>
<td>$X_5$</td>
<td>$p(x_5^j</td>
</tr>
<tr>
<td>$X_6$</td>
<td>$p(x_6^j</td>
</tr>
<tr>
<td>$X_7$</td>
<td>$p(x_7^j</td>
</tr>
<tr>
<td>$X_8$</td>
<td>$p(x_8^j</td>
</tr>
</tbody>
</table>
Experiments

Mass spectrometry datasets

<table>
<thead>
<tr>
<th>Name</th>
<th># C1</th>
<th># C2</th>
<th>Size</th>
<th># Features</th>
<th>SDR</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ovarian cancer profiling</td>
<td>121</td>
<td>79</td>
<td>200</td>
<td>45,200</td>
<td>0,0044</td>
<td>Petricoin et al. (2002)</td>
</tr>
<tr>
<td>Detection of drug-induced toxicity</td>
<td>28</td>
<td>34</td>
<td>62</td>
<td>45,200</td>
<td>0,00137</td>
<td>Petricoin et al. (2004)</td>
</tr>
<tr>
<td>Hepatocellular carcinoma</td>
<td>78</td>
<td>72</td>
<td>150</td>
<td>36,802</td>
<td>0,0041</td>
<td>Ressom et al. (2006)</td>
</tr>
</tbody>
</table>

- Estimation algorithms: UMDA, BMDA
- Classifiers: Naive Bayes, k-NN, SVM
- Average all EDA results over 500 multistarts
Results [preliminary]

Usage for knowledge discovery: peak frequency plots
Future challenges

- Better dealing with correlated features
  - First cluster correlated features, then choose representatives from each cluster, and build a model with the representatives
  - Adapt similarity measures to deal with correlated features
Future challenges

- Better dealing with correlated features
  - First cluster correlated features, then choose representatives from each cluster, and build a model with the representatives
  - Adapt similarity measures to deal with correlated features

- Increasing stability by transfer learning
  - Assume 2 related datasets $D_1$ and $D_2$
  - Use feature selection on $D_1$ as “prior” for feature selection on $D_2$
  - Preliminary research shows that this “transferral” of feature selection information increases the stability of feature selection on $D_2$ [Helleputte and Dupont (2009)]
Future challenges

- Better dealing with correlated features
  - First cluster correlated features, then choose representatives from each cluster, and build a model with the representatives
  - Adapt similarity measures to deal with correlated features

- Increasing stability by transfer learning
  - Assume 2 related datasets $D_1$ and $D_2$
  - Use feature selection on $D_1$ as “prior” for feature selection on $D_2$
  - Preliminary research shows that this “transferral” of feature selection information increases the stability of feature selection on $D_2$ [Helleputte and Dupont (2009)]

- A comparative evaluation of different ensemble FS techniques
Acknowledgements

- Thomas Abeel (Ghent University)
- Yves Van de Peer (Ghent University)
- Thibault Helleputte (UC Louvain)
- Pierre Dupont (UC Louvain)
- Ruben Armañanzas (Universidad Politecnica de Madrid)
- Iñaki Inza (University of the Basque Country)
- Pedro Larrañaga (Universidad Politecnica de Madrid)


