NIPS 2003 Feature Selection Competition

Yi-Wei Chen and Chih-Jen Lin

Department of Computer Science
National Taiwan University

NIPS 2003, December, 2003
Choosing Tools in the Beginning

• Simple statistical measures
  F scores

• Classification methods:
  – Support vector machines (SVM)
  – Random forest

• Reasons:
  We are more familiar with these two methods
  They are rather simple
SVM Feature Selection

- Direct use without feature selection
  Sometimes good enough
- SVM with linear kernel, choose larger primal coefficients
  Not considered here
- Radius margin bound with RBF kernel:
  Modified RBF kernel
  \[ K(x, y) = \exp(-g_1(x_1 - y_1)^2 - \cdots - g_n(x_n - y_n)^2) \]
Minimize leave-one-out (loo) bound:

$$\text{loo} \leq f(C, g_1, \ldots, g_n)$$

- $g_i$ close to zero, less important

Two-level minimization:
$C, g_1, \ldots, g_n$ fixed: SVM optimization problem
if $f$ carefully constructed, it is differentiable
But still difficult non-convex problems, $n$ cannot be too large
Random Forest Feature Selection

- 500 trees
  Each tree: using a fixed number of random features
- Each tree: out of bag validation
  Feature importance
SVM and Random Forest

• Our experience:
  Same data, with full parameter selection
  SVM slightly better than RM

• But SVM requires higher cost on training+parameter selection
  SVM more sensitive to parameters

• Random Forest directly gives feature importance
  Mainly used here for selecting features
  i.e., after features selected, still use SVM for prediction
## Things We Have Tried

- **Validation error:**

<table>
<thead>
<tr>
<th>Method</th>
<th>arcene</th>
<th>dexter</th>
<th>dorothea</th>
<th>gisette</th>
<th>madelon</th>
</tr>
</thead>
<tbody>
<tr>
<td>simple SVM</td>
<td>0.1331</td>
<td>0.1167</td>
<td>0.3398</td>
<td>0.0210</td>
<td>0.4017</td>
</tr>
<tr>
<td>F + SVM</td>
<td>0.2143</td>
<td>0.0800</td>
<td>0.2138</td>
<td>0.0180</td>
<td>0.1300</td>
</tr>
<tr>
<td>F + RF + SVM</td>
<td>0.3295</td>
<td>0.0867</td>
<td>0.1251</td>
<td>0.0400</td>
<td>0.0767</td>
</tr>
<tr>
<td>RF + RM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.0750</td>
</tr>
<tr>
<td>F + RF + RM</td>
<td></td>
<td>0.1430</td>
<td></td>
<td></td>
<td>0.0850</td>
</tr>
</tbody>
</table>

- **F:** F score; **RF:** Random Forest
- **SVM:** Support vector machines
- **RM:** radius margin bound
• We focus more on the first three approaches

• Each attribute scaled to [0,1] first

• F score: threshold determined by either CV or human eyes

<table>
<thead>
<tr>
<th></th>
<th>arcene</th>
<th>dexter</th>
<th>dorothea</th>
<th>gisette</th>
<th>madelon</th>
</tr>
</thead>
<tbody>
<tr>
<td>threshold</td>
<td>0.1</td>
<td>0.015</td>
<td>0.05</td>
<td>0.01</td>
<td>0.005</td>
</tr>
</tbody>
</table>

• After selecting features, parameter selection on training set conducted (with RBF kernel)
Final Submission

- Using those with the smallest validation error

<table>
<thead>
<tr>
<th></th>
<th>train error</th>
<th>valid error</th>
<th>test error</th>
<th>#features</th>
</tr>
</thead>
<tbody>
<tr>
<td>arcene</td>
<td>0.0000</td>
<td>0.1331</td>
<td>0.1527</td>
<td>10000 (100%)</td>
</tr>
<tr>
<td>dexter</td>
<td>0.0033</td>
<td>0.0800</td>
<td>N/A</td>
<td>209 (1.04%)</td>
</tr>
<tr>
<td>dorothea</td>
<td>0.0256</td>
<td>0.1251</td>
<td>N/A</td>
<td>445 (0.45%)</td>
</tr>
<tr>
<td>gisette</td>
<td>0.0000</td>
<td>0.0180</td>
<td>0.0137</td>
<td>913 (18.26%)</td>
</tr>
<tr>
<td>madelon</td>
<td>0.0370</td>
<td>0.0750</td>
<td>0.0661</td>
<td>24 (4.8%)</td>
</tr>
</tbody>
</table>

- test error: December 1

- final1 and final2: the same thing except arcene
  a mistake in final1 for arcene
Discussion: SVM and gisette

- gisette: modified from MNIST digit recognition
  Simple SVM works well for this problem

<table>
<thead>
<tr>
<th></th>
<th>simple SVM</th>
<th>F + SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>validation error</td>
<td>0.0210</td>
<td>0.0180</td>
</tr>
</tbody>
</table>

- SVM’s problem when # features large:
  RBF kernel

\[ K(x, y) = e^{-g\|x-y\|^2} \]

  Same \( g \) for relevant and irrelevant features

- My experience on MNIST (784 features) and USPS (256 features):
  Features from the same kind of “sources”: this issue less serious
  larger #features can be handled.
• Additional features generated from “products of pairs of variables”
  Probes: similar distribution
  This may be why SVM without feature selection works well

• Another problem simple SVM works well is arcene
  Reason?
Discussion: Radius Margin Bound and Madlon

- The only problem that we find RM bound useful
- Good results by Wei Chu
  
  I guess they use Bayesian SVM [Chu, Keerthi, Ong]
  
  Under Bayesian framework,

  $$\min f(C, g_1, \ldots, g_n)$$

- Though two different derivations
  
  Formula a little bit related to the RM loo bound

- In practice: once Keerthi told me that when testing some UCI problems, Bayesian SVM works similar to using one single $g$, but improve 5% on splice
  
  We then checked the RM bound
The same result

- Looks like this problem is another *splice*

- Issue: Can we know from the generation of this data why the two formulas work?
Conclusions

• The whole procedure a bit ad hoc
  More systematic procedures?
• Domain knowledge not used
• We thank organizers for this interesting competition