Evaluation of kernel function modification in text classification using SVMs

Yangzhe Xiao
Motivation

• One of the limitations of SVMs lies in the choice of kernels.

• Good choices of kernels can make a linearly inseparable case become separable in feature space or increase the margin in feature space.
Kernel Modification

- Christianini, Shawe-Taylor & Lodhi (2001) used Latent Semantic Indexing (LSI) to extract the semantic relation matrix $M$, and modified radial basis kernel (RBF) as
  
  \[ K(x,y)=\exp\left(-\frac{||Mx-My||^2}{2\sigma^2}\right). \]

--Documents are implicitly mapped into a “semantic space”.
--Better results.

This study is motivated by the same modification idea but with different $M$. 
Measurement of Discriminating Power

• For a term $t$ and a category $c$,

<table>
<thead>
<tr>
<th></th>
<th>$c$</th>
<th>Non-$c$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$t$</td>
<td>$A$</td>
<td>$B$</td>
</tr>
<tr>
<td>No-$t$</td>
<td>$C$</td>
<td>$D$</td>
</tr>
</tbody>
</table>

\[
\chi^2(t, c) = \frac{N \times (AD - CB)^2}{(A + C) \times (B + D) \times (A + B) \times (C + D)}
\]

$N = A + B + C + D$.

$\chi^2(t,c)$ is zero if $t$ and $c$ are independent.
Measurement of Discriminating Power (cont’d)

- For a term $t$ and a category $c$,

<table>
<thead>
<tr>
<th></th>
<th>$c$</th>
<th>Non-$c$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$t$</td>
<td>$A$</td>
<td>$B$</td>
</tr>
<tr>
<td>No-$t$</td>
<td>$C$</td>
<td>$D$</td>
</tr>
</tbody>
</table>

$$E(t, c) = \frac{A}{A + C} \left(1 - (p \log_2 p - q \log_2 q)\right)$$

where

$$p = \frac{\frac{A}{A + C} + \frac{B}{B + D}}{\frac{A}{A + C} + \frac{C}{A + C} + \frac{B}{B + D}}$$

$q = 1 - p$
Matrix $M$

- Assign weights for terms according to the two measurements $\chi^2(t,c)$ or $E(t,c)$ by modifying kernel.
- Let $M$ be a square diagonal matrix with values of $(1+\chi^2(t,c))$ or $(1+E(t,c))$ for every terms in the feature set.
Documents

• Training and testing documents are from Reuters-21578 and have 10 categories.

• Precision ($\pi$) and recall ($\rho$) are defined as:

$$\pi_i = \frac{TP_i}{TP_i + FP_i} \quad \rho_i = \frac{TP_i}{TP_i + FN_i}.$$

Here, $FP_i$ (false positives wrt category $c_i$) is the number of test documents incorrectly classified under $c_i$; $TN_i$ (true negatives wrt $c_i$), $TP_i$ (true positives wrt $c_i$), and $FN_i$ (false negative wrt $c_i$) are defined accordingly.
Results

<table>
<thead>
<tr>
<th>Sample Size</th>
<th>T</th>
<th>CW</th>
<th>EW</th>
<th>TS</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>0.185</td>
<td>0.211</td>
<td>0.223</td>
<td>0.197</td>
</tr>
<tr>
<td>100</td>
<td>0.574</td>
<td>0.602</td>
<td>0.613</td>
<td>0.593</td>
</tr>
<tr>
<td>250</td>
<td>0.690</td>
<td>0.726</td>
<td>0.728</td>
<td>0.717</td>
</tr>
<tr>
<td>500</td>
<td>0.779</td>
<td>0.790</td>
<td>0.791</td>
<td>0.788</td>
</tr>
<tr>
<td>1000</td>
<td>0.810</td>
<td>0.819</td>
<td>0.820</td>
<td>0.814</td>
</tr>
<tr>
<td>1880</td>
<td>0.837</td>
<td>0.855</td>
<td>0.851</td>
<td>0.855</td>
</tr>
<tr>
<td>5100</td>
<td>0.830</td>
<td>0.861</td>
<td>0.867</td>
<td>0.880</td>
</tr>
</tbody>
</table>

\[
F = \frac{2 \pi \rho}{\pi + \rho}
\]

\(F\) depends equally on precision \(\pi\) and recall \(\rho\).
The End