Feature selection methods are essential for learning to rank (LTR) approaches as the number of features are directly proportional to computational cost and sometimes, might lead to the over-fitting of the ranking model. We propose an expected divergence based approach to select a subset of highly discriminating features over relevance categories.

FS-ED: Method Details

The proposed method has two components:

(i) Importance of the features \( s(f_i) \): The evaluation measure, NDCG@10, to estimate the importance of an individual feature and,

(ii) Expected divergence of the features \( d(f_i) \): The divergence of the features over relevance classes taking ordinal class information into account.

\[
d(f_i) = \sum_{m=0}^{R_i-1} \sum_{n=m+1}^{R_i-1} (n-m) \cdot \text{div}(f_i^{(n)}, f_i^{(m)})
\]

where, \( \text{div}(f_i^{(n)}, f_i^{(m)}) = \frac{1}{2}d_{KL}(f_i^{(n)}||f_i^{(m)}) + \frac{1}{2}d_{KL}(f_i^{(m)}||f_i^{(n)}) \)

The goal of the method is to score each feature \( f_i \in F \), where \( F \) is the set of all features and \( |F| = n \). We pose the feature selection method as a maximization problem of selecting top \( k \) features from \( F \) where, the score of a feature \( \psi(f) \) is calculated as shown in Eq. 1. For the simplicity, we combine the two objective functions linearly.

\[
\psi(f_i) = s(f_i) + d(f_i)
\]

Algorithm

\[ T = \text{training data} \]
\[ V = \text{validation data} \]
\[ \psi = \text{weight vector of features} \]
\[ F, F_k = \text{feature sets, all and top-k respectively} \]

for each \( f_i \in F \)

\[ \psi(f_i) = 0 \quad / \ast \text{Initialise the weights} / \ast \]

end for

for each \( f_i \)

\[ s(f_i) = \text{evaluation score over \( T \)} \]

for each \( r_i \in R \)

\[ \text{estimate PDF}(f_i) \] over \( r_i \) from \( T \) using KDE \]

end for

for \( i = 0 \) to \( |R| - 1 \)

for \( j = i + 1 \) to \( |R| - 1 \)

\[ \text{estimate JS div. of} f_i \text{ over} r_i \text{ and} r_j \text{ from} V \]

end for

end for

\[ \psi(f_i) = s(f_i) + d(f_i) \]

end for

\[ \text{sort} \; \psi \]

for \( i = 1 \) to \( k \)

\[ \text{add} \; f_i \text{ in} \; F_k \]

end for

RETURN \( F_k \)

Expected Divergence based Feature Selection for Learning to Rank

Parth Gupta and Paolo Rosso

http://www.dsic.upv.es/grupos/nle {pgupta,prosso}@dsic.upv.es

Contribution: Fast and Scalable Feature Selection Method

Experimental Setup

In order to compare the proposed method with the baselines, we use the performance evaluation of three state-of-the-art LTR algorithms when trained with selected features on four standard LTR datasets. We use NDCG@10 as metric and 5-fold cross validation.

Baseline 1: FS-BFS The FS-BFS is a wrapper based approach of feature selection for ranking [Dang and Croft, 2010]. The method partitions the \( F \) into non-overlapping \( k \) subsets and learns a ranking model which maximizes the performance over that subset of features. Best first search is used on the undirected graph of features to extract subsets and the weights of the features are learnt using coordinate ascent.

Baseline 2: WAF With all Features.

Analysis

It is noticeable that, the top features better discriminate between the relevance classes and exhibit high divergence over distant relevance levels.

Results

The number of features used to obtain the reported results with different feature selection strategies. O, M7, M8 and Y refer to OHSUMED, MQ2007, MQ2008 and Yahoo! respectively.

<table>
<thead>
<tr>
<th>FS Method</th>
<th>O</th>
<th>M7</th>
<th>M8</th>
<th>Y</th>
<th>R. Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>FS-ED</td>
<td>4</td>
<td>15</td>
<td>15</td>
<td>75</td>
<td>RankSVM</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>20</td>
<td>10</td>
<td>75</td>
<td>RankBoost</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>14</td>
<td>10</td>
<td>7</td>
<td>LambdaMART</td>
</tr>
<tr>
<td>FS-BFS</td>
<td>6</td>
<td>9</td>
<td>12</td>
<td>-</td>
<td>RankSVM</td>
</tr>
<tr>
<td></td>
<td>12</td>
<td>7</td>
<td>-</td>
<td>-</td>
<td>RankBoost</td>
</tr>
<tr>
<td></td>
<td>14</td>
<td>10</td>
<td>-</td>
<td>-</td>
<td>LambdaMART</td>
</tr>
<tr>
<td>WAF</td>
<td>45</td>
<td>46</td>
<td>46</td>
<td>699</td>
<td>ALL</td>
</tr>
</tbody>
</table>

References


Remarks

- As the score calculation of a feature does not depend on other features’ scores, hence can be parallelised.
- The proposed method leads to not significantly worse, and in some cases, significantly better performance compared to the baselines.
- Desired results are achieved with as few features as less than 10% on a set of standard datasets and state-of-the-art LTR algorithms.
- Analysis exhibit that large margin classifier based ranking models can greatly benefit from this selection method.