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

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.

$$d(f_i) = \sum_{m=0}^{|R|-1} \sum_{n=m+1}^{|R|-1} (n-m) * \text{div}(f_i^{r_m}, f_i^{r_n})$$

where, $\text{div}(f_i^{r_m}, f_i^{r_n}) =$

$$\frac{1}{2} d_{KL}\left(\hat{f}_i^{r_m} || \hat{f}_i^{avg}\right) + \frac{1}{2} d_{KL}\left(\hat{f}_i^{r_n} || \hat{f}_i^{avg}\right)$$

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 $\psi(\cdot)$ is calculated as shown in Eq. 1. For the simplicity, we combine the two objective functions linearly.

$$\psi(f_i) = s(f_i) + d(f_i) \qquad (1)$$

Algorithm

```
$T$ = training data
$V$ = validation data
$\vec{\psi}$ = weight vector of features
$F, F_k$ = feature sets, all and top-$k$ respectively
for each $f_i \in F$
    $\psi(f_i) = 0$    /* Initialise the weights */
end for
for each $f_i$
    $s(f_i)$ = evaluation score over $T$
    for each $r_i \in R$
      estimate $PDF(f_i)$ over $r_i$ from $T$ using KDE
    end for
    for $i = 0$ to $|R| - 1$
      for $j = i + 1$ to $|R| - 1$
        estimate JS div. of $f_i$ over $r_i$ and $r_j$ from $V$
      end for
    end for
    calculate $d(f_i)$ as show in Eq. 1
    $\psi(f_i) = s(f_i) + d(f_i)$
end for
sort $\vec{\psi}$
for $i = 1$ to $k$
    add $f_i$ in $F_k$
end for
RETURN $F_k$
```

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**.

Experimental Setup

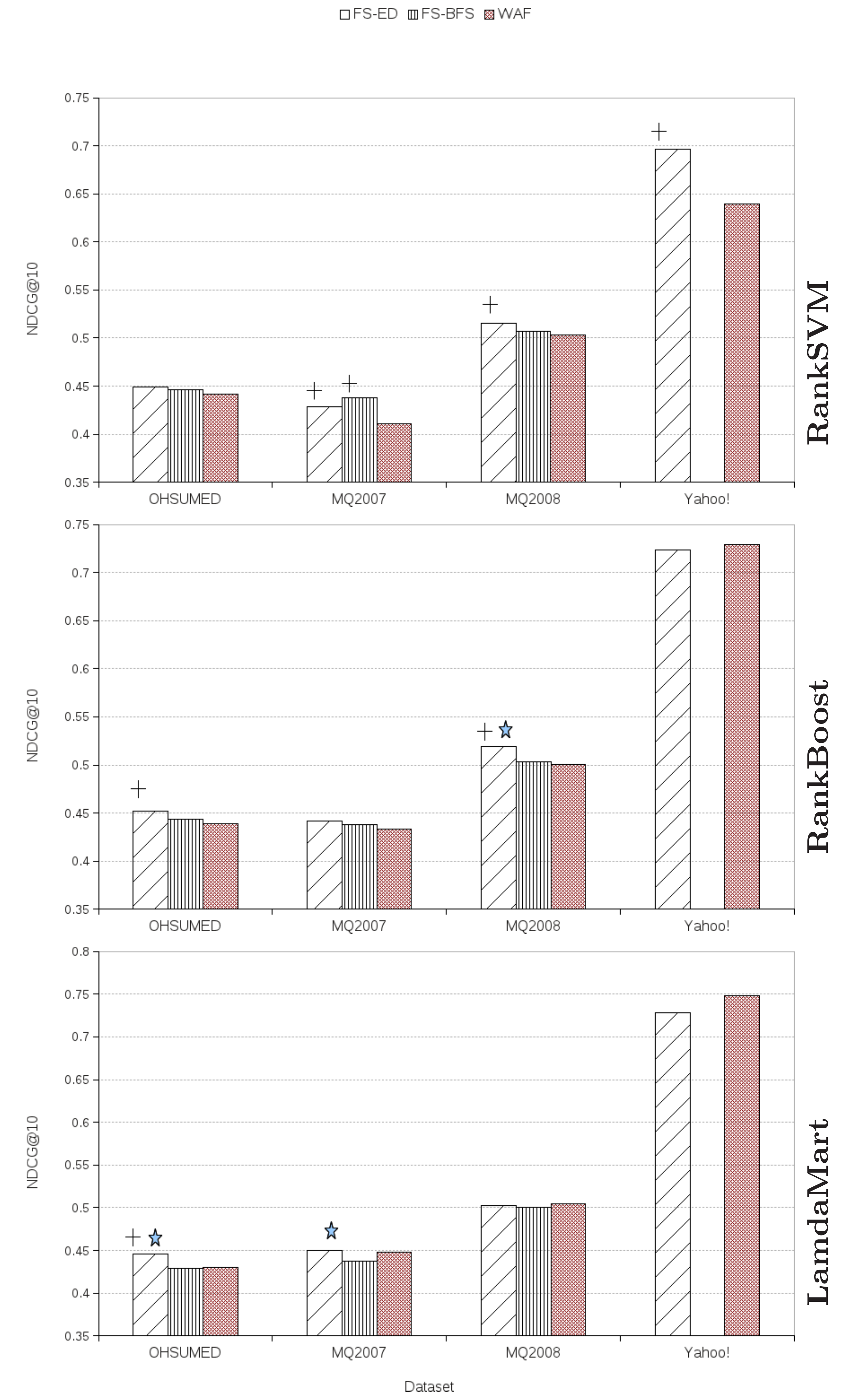
In order to compare the proposed method with the baselines, we use the performance evaluation of three state-of-the-art LTR algorithms when trained with selected features on four standard LTR datasets. We use NDCG@10 as metric and 5-fold cross validation.

***(i)* RankSVM** SVM based pairwise ranker.
***(ii)* RankBoost** Weak ranker based pairwise ranker that uses boosting.
***(iii)* LambdaMART** LambdaMART uses gradient boosting to optimize a ranking cost function.

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.

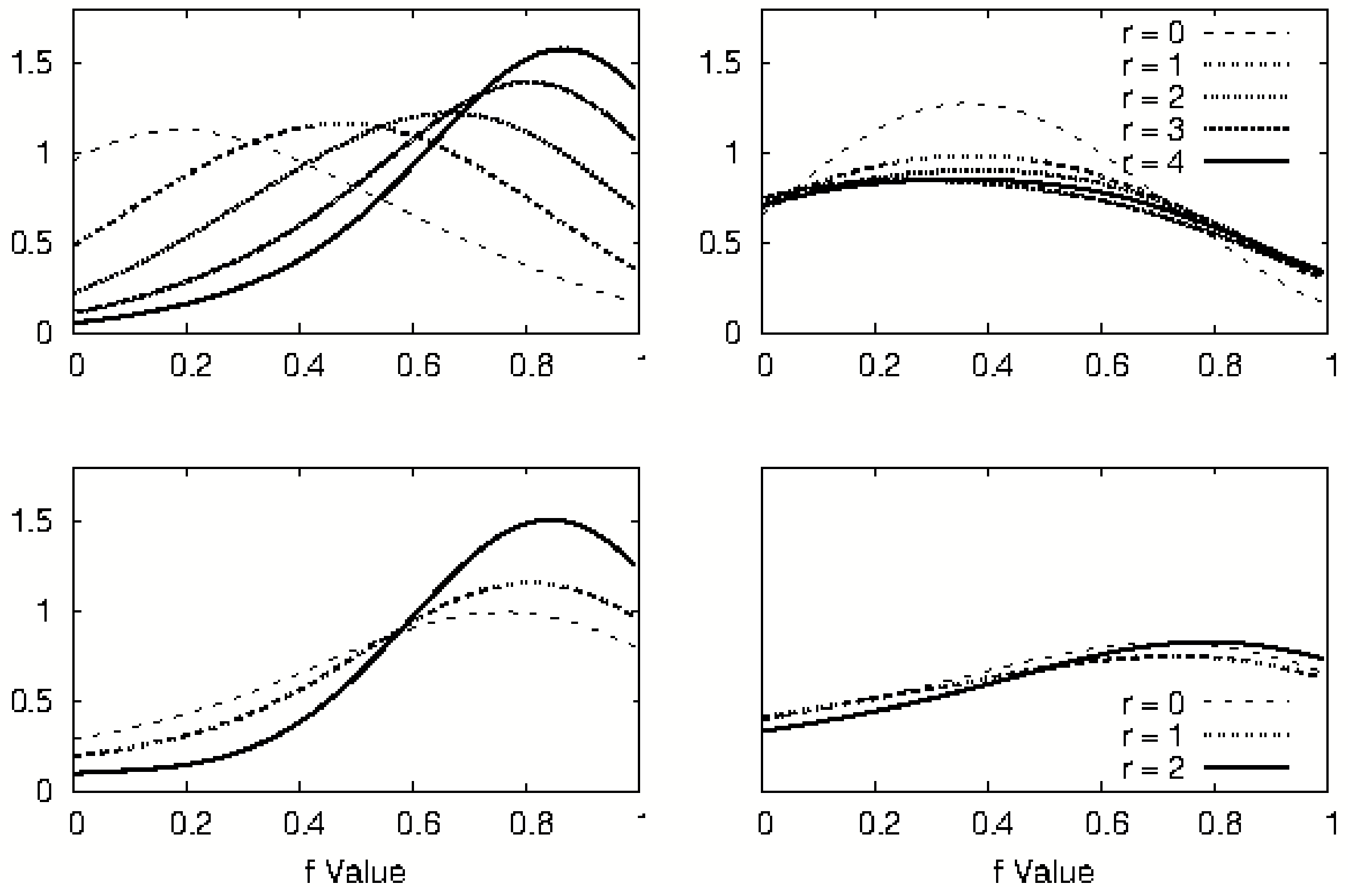
Results



+ and ⋆ indicate statistical significance with WAF and the other FS strategy respectively.

Analysis

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



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

FS Method	O	M7	M8	Y	R. Method
FS-ED	15	**3**	**3**	**50**	RankSVM
	4	15	15	75	RankBoost
	20	10	20	75	LambdaMART
FS-BFS	**6**	9	12	-	RankSVM
	12	**7**	**7**	-	RankBoost
	14	10	**7**	-	LambdaMART
WAF	45	46	46	699	ALL

References

[Dang and Croft, 2010] Dang, V. and Croft, B. W. (2010). Feature selection for document ranking using best first search and coordinate ascent. In *SIGIR Workshop on Feature Generation and Selection for Information Retrieval*,, SIGIR '10.