

# Evaluating Feature Selection Methods for Multi-Label Text Classification

Newton Spolaôr<sup>1</sup>, Grigorios Tsoumakas<sup>2</sup>

<sup>1</sup> Laboratory of Computational Intelligence,  
Institute of Mathematics & Computer Science  
University of São Paulo, São Carlos, Brazil

<sup>2</sup> Department of Informatics  
Aristotle University of Thessaloniki  
Thessaloniki 54124, Greece



# Motivation

- Real world, exciting research problem on large-scale biomedical semantic indexing



- Can feature selection help?

# Multi-Label Learning

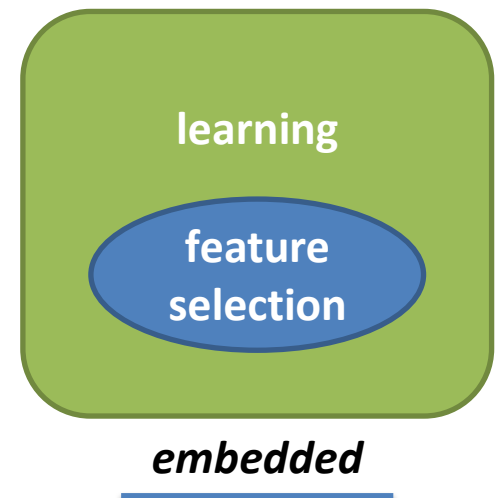
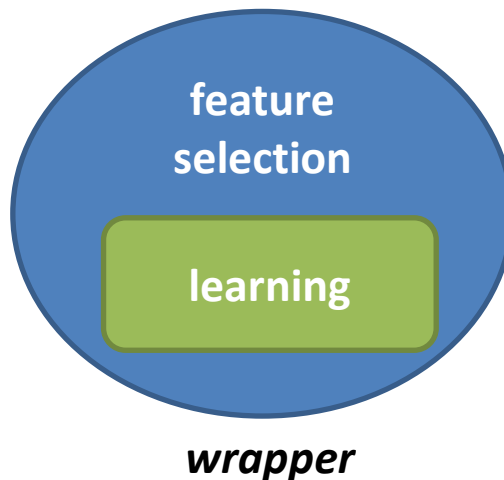
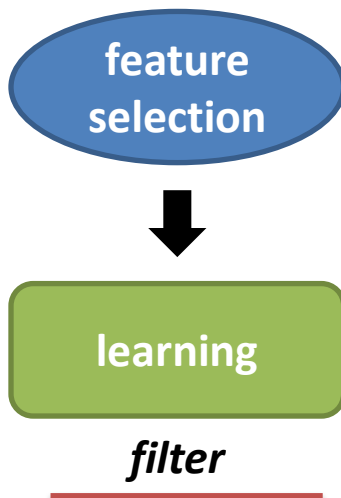
- Multi-label data
  - Instances related with a subset of a finite label set

	Pain	Fever	...	Weight	Disease
Patient 1	yes	no		101.5	{gastritis, duodenitis}
Patient 2	no	yes		61.2	{esophagitis}
⋮	⋮	⋮	⋮	⋮	⋮
Patient M	yes	yes	...	79.8	{esophagitis, gastritis, duodenitis}

- Models learned from such data can output
  - Bipartition of label set, ranking of labels, ranking of instances, marginal/joined probabilities

# Feature Selection



- Main objectives
  - Reducing measurement & storage requirements, data understanding, reducing training & utilization times, improving prediction accuracy
- Three main categories of approaches



# Multi-Label Filter Feature Selection


- Step 1: Feature ranking separately per label
  - One can use any standard single-label feature evaluation measure for binary classification
- Step 2: Aggregation of the different rankings
  - Mean, Max of the evaluation score for all labels
  - Round Robin (RoR), Rand Robin (RaR) selection per label based on the evaluation scores

# Example – Mean Aggregation


Feature	Score $Y_1$	Score $Y_2$	Score $Y_3$		Mean		Ranking
$X_1$	0.1	0.9	0.5		0.5		$X_5$
$X_2$	0.6	0	0.3		0.3		$X_3$
$X_3$	0.5	0.7	0.6		0.6		$X_1$
$X_4$	0.3	0.5	0.4		0.4		$X_4$
$X_5$	0.7	0.6	0.8		0.7		$X_2$

# Example – Max Aggregation

Feature	Score $Y_1$	Score $Y_2$	Score $Y_3$
$X_1$	0.1	0.9	0.5
$X_2$	0.6	0	0.3
$X_3$	0.5	0.7	0.6
$X_4$	0.3	0.5	0.4
$X_5$	0.7	0.6	0.8



Max
0.9
0.6
0.7
0.5
0.8



Ranking	Mean
$X_1$	$X_5$
$X_5$	$X_3$
$X_3$	$X_1$
$X_2$	$X_4$
$X_4$	$X_2$


# Example – RoR Aggregation

Feature	Score $Y_1$	Score $Y_2$	Score $Y_3$		Ranking
$X_1$	0.1	0.9	0.5		$X_5$
$X_2$	0.6	0	0.3		$X_1$
$X_3$	0.5	0.7	0.6		$X_2$
$X_4$	0.3	0.5	0.4		$X_3$
$X_5$	0.7	0.6	0.8		$X_4$

Diagram illustrating the RoR Aggregation process. The input table shows scores for five features ( $X_1$  to  $X_5$ ) across three criteria ( $Y_1$ ,  $Y_2$ ,  $Y_3$ ). The scores are:  $X_1$  (0.1, 0.9, 0.5),  $X_2$  (0.6, 0, 0.3),  $X_3$  (0.5, 0.7, 0.6),  $X_4$  (0.3, 0.5, 0.4), and  $X_5$  (0.7, 0.6, 0.8). The scores 0.9, 0.6, 0.7, 0.7, and 0.8 are circled in red. Red arrows point to the columns: two arrows for  $Y_1$ , two for  $Y_2$ , and one for  $Y_3$ . A blue arrow points from the input table to the resulting ranking table, which lists the features in descending order of their aggregated scores:  $X_5$ ,  $X_1$ ,  $X_2$ ,  $X_3$ , and  $X_4$ .



# Example – RaR Aggregation



Feature	Score $Y_1$	Score $Y_2$	Score $Y_3$
$X_1$	0.1	0.9	0.5
$X_2$	0.6	0	0.3
$X_3$	0.5	0.7	0.6
$X_4$	0.3	0.5	0.4
$X_5$	0.7	0.6	0.8
<b>frequency</b>	<b>0.3</b>	<b>0.1</b>	<b>0.4</b>



Ranking	RoR
$X_5$	$X_5$
$X_3$	$X_1$
$X_1$	$X_2$
$X_2$	$X_3$
$X_4$	$X_4$

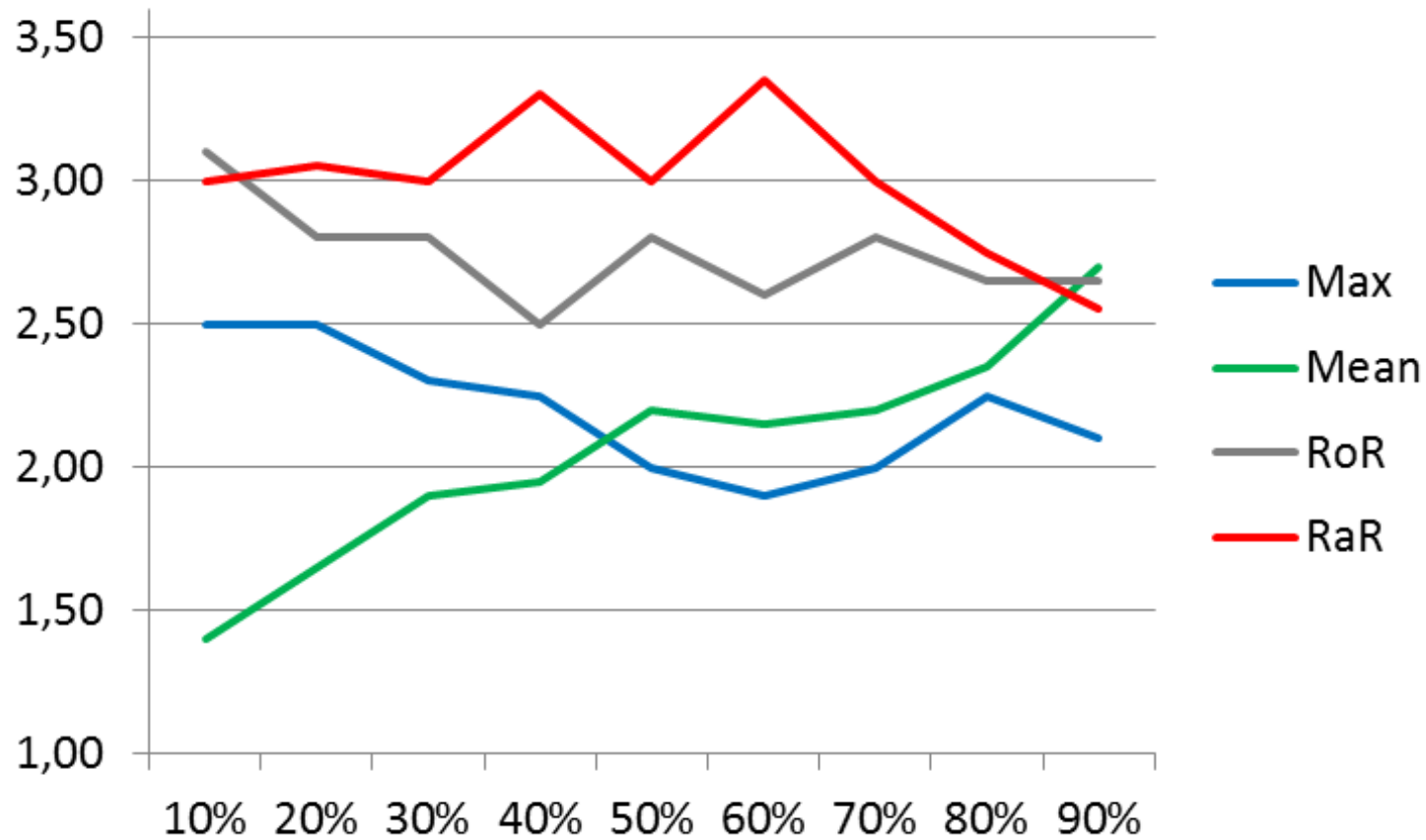
# Experimental Setup (1/2)

- 20 benchmark textual datasets
  - yahoo (11), enron, delicious, bookmarks, bibtex, medical, tmc007, slashdot, language log, rcv1v2
- 8 filter feature selection methods
  - 2 feature evaluation measures ( $\chi^2$ , BNS)
  - 4 aggregation strategies (Mean, Max, RoR, RaR)
- 2 baselines
  - random feature selection (RFS), all features (AF)

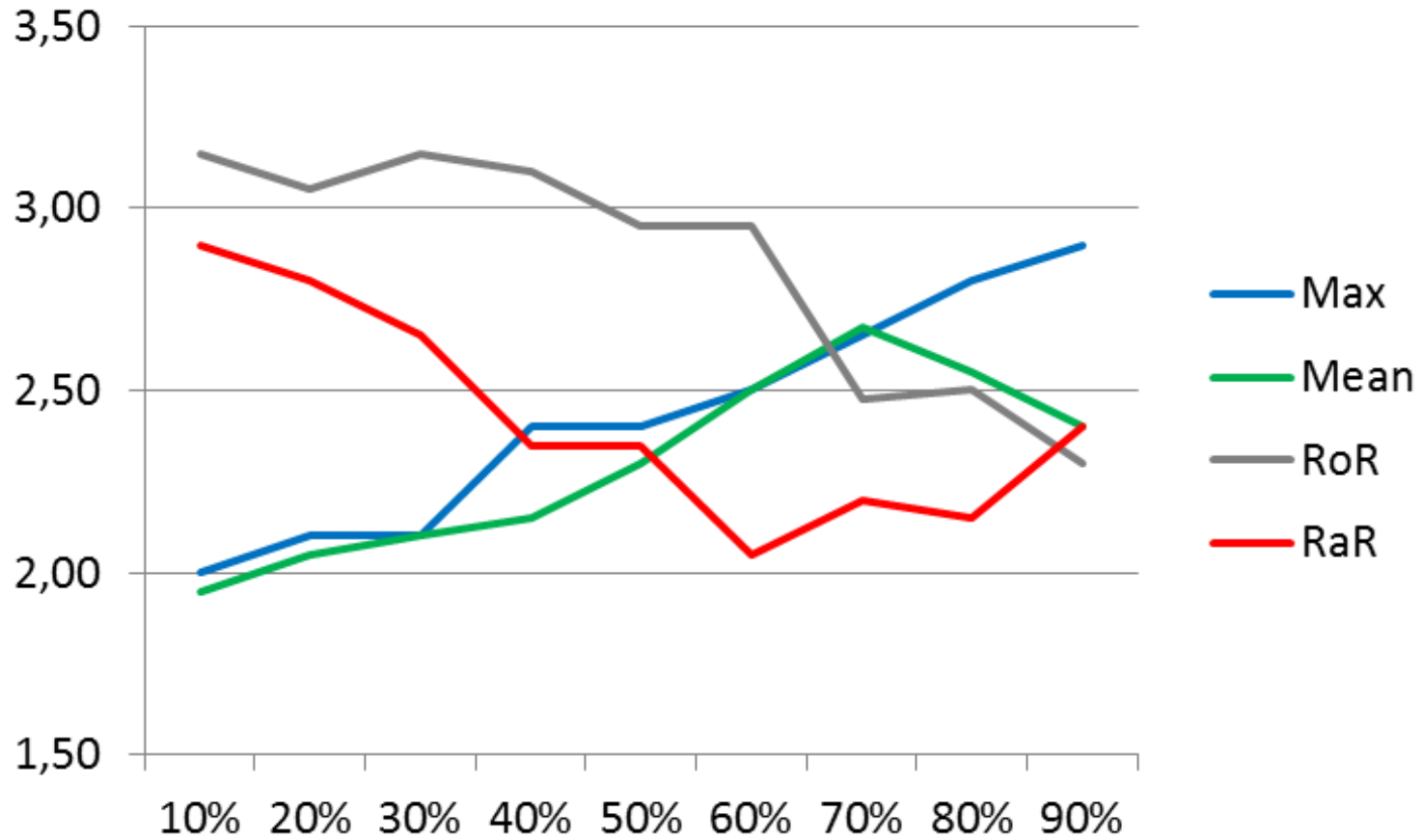
# Experimental Setup (2/2)

- Multi-label classification
  - Binary Relevance (*aka one-vs-rest*) with linear support vector machines as base algorithm
- Evaluation
  - Micro F-measure
  - Selection of 10%, 20%, ..., 90% features
  - Average ranking of methods across datasets

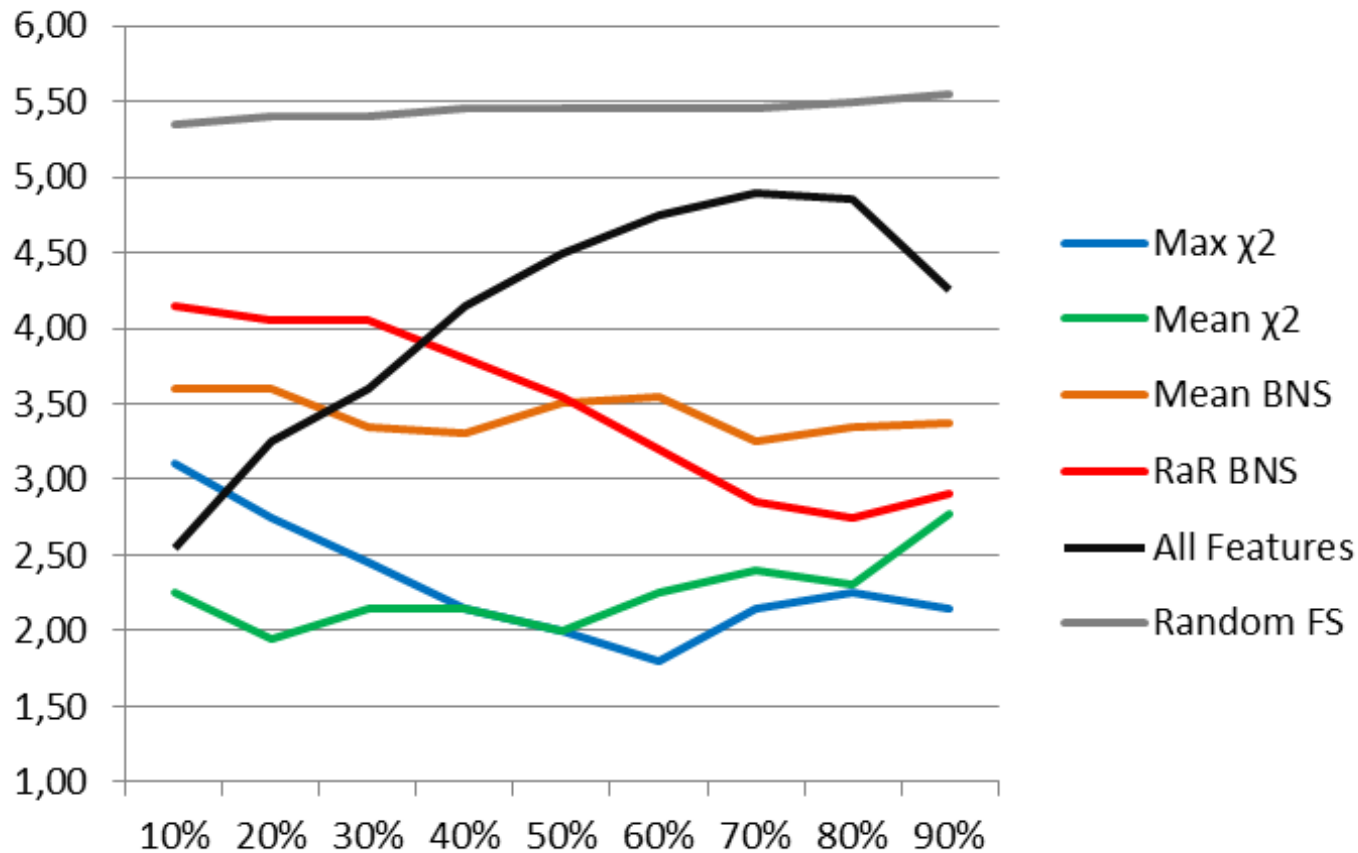
# $\chi^2$ Results



# BNS Results



# Best of $\chi^2$ and BNS, Random, All



# Recap

- Empirical study with large number of text datasets (20) in contrast with past literature
- Aggregations RaR and RoR tried for the first time here, but did not work successfully
- BNS is worse than  $\chi^2$ , contrary to findings for single-label data
- For  $\chi^2$  mean (max) aggregation should be preferred for low (high) percentage of features

# Future Work

- Binary relevance + global feature selection
- Binary relevance + local feature selection
- Meta-labeler + global feature selection
  - First results on BioASQ data are negative
  - Will verify this on the 20 datasets of this study
- Meta-labeler + local feature selection
  - Fails, as it renders the SVM scores incomparable
- Explore efficient ways to exploit label dependence in multi-label feature selection



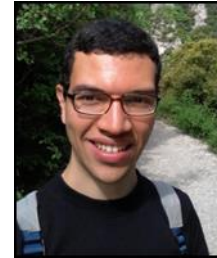
# The End

- Thank you for your attention!

- Contact

- [newtonspolaor@gmail.com](mailto:newtonspolaor@gmail.com)

- [greg@csd.auth.gr](mailto:greg@csd.auth.gr)



- Acknowledgement

- This research was partially supported by the São Paulo Research Foundation (FAPESP), grant 2012/23906-2

