WISE 2014 Challenge - 7th place

Eleftherios Spyromitros-Xioufis, PhD student, MLKD group,
Department of Informatics, Aristotle University of Thessaloniki

Main problem characteristics

- Each document can be assigned to more than one category
  - → Use **M**ulti-**L**abel **C**lassification (**MLC**) algorithms
- Documents given as tf-idf vectors
  - → Focus on the learning problem instead of the representation
- Large dimensionality (features, labels, examples)
  - → Seek efficient solutions
- Train/test documents are chronologically ordered
  - → Possible gains by exploit this info

Eleftherios Spyromitros-Xioufis, PhD Student
espyromi@csd.auth.gr

Evaluation methodology

- Reliable internal evaluation is essential
  - Estimated performance should reflect test performance
  - Allows improving the model without needing LB feedback
- A simple but effective recipe: mimic the given train/test split
  - First 65% for training, last 35% for testing
  - Assumes documents are chronologically ordered within train set
- Internal evaluation results correlate very well with LB results

Run pos.	$F_1^{internal}$	F_1^{public}	$F_1^{private}$
1	0.7806	0.7793	0.7827
2	0.7806	0.7789	0.7819
3	0.7794	0.7788	0.7819

Eleftherios Spyromitros-Xioufis, PhD Student
espyromi@csd.auth.gr

Main ingredients of the solution

- A plug-in rule approach for F_1 maximization
  - Probabilistic ML classifier → Binary Relevance + Logistic Regression
  - Derive F_1 optimal predictions during inference → ML^E approach [1]
- Feature selection using χ^2_{max} [2]
  - 4% improvement over using all features
  - A model using 6% of the features in the top 10
- A normalization of the features, inspired from Computer Vision
  - 0.3% improvement
- A multi-view ensemble scheme
  - Averages outputs of multiple models built on different top-k subsets
  - 1.2% improvement over the best standalone model

[1] Ye et al. Optimizing F-measures: A Tale of Two Approaches. ICML 2012.
[2] Lewis et al. Rcv1: A new benchmark collection for text categorization research. JMLR 2004.

Eleftherios Spyromitros-Xioufis, PhD Student
espyromi@csd.auth.gr

F_1 maximization

- Two strategies
  - Structured loss minimization:
    - optimizes F_1 directly during training (e.g. SSVM)
  - Plug-in rule:
    - a probabilistic model +
    - a separate inference step to derive optimal predictions
- Why a plug-in rule method?
  - Better with rare classes [1] (very common in MLC problems)
  - Better in MLC experiments with example-based F_1 [3]
  - More efficient during training [3]

[1] Ye et al. Optimizing F-measures: A Tale of Two Approaches. ICML 2012.
[3] Dembczynski et al. Optimizing the F-Measure in Multi-label Classification: Plug-in Rule Approach versus Structured Loss Minimization. ICML 2013.

Eleftherios Spyromitros-Xioufis, PhD Student
espyromi@csd.auth.gr

Plug-in rule realization

- The ML^E approach [1]
  - Originally proposed in a binary classification context
  - Exploits the *Probability Ranking Principle* (*PRP*) [2]
  - *PRP*: Under the assumption of independence solution contains the k most likely labels
  - *Computes F_β optimal predictions in $O(n^2)$* for reasonable values of β
- Algorithms that do not make the independence assumption [3]
  - Similar results in practice
  - Significantly slower
- ML^E needs a MLC model providing good marginals
  - Binary Relevance + probabilistic binary classifier

[1] Ye et al. Optimizing F-measures: A Tale of Two Approaches. ICML 2012.
[2] Lewis et al. Rcv1: A new benchmark collection for text categorization research. JMLR 2004.
[3] Dembczynski et al. Optimizing the F-Measure in Multi-label Classification: Plug-in Rule Approach versus Structured Loss Minimization. ICML 2013.

Eleftherios Spyromitros-Xioufis, PhD Student
espyromi@csd.auth.gr

Binary Relevance

- One-vs-all or Binary Relevance (BR): $h(\boldsymbol{x}) \rightarrow \boldsymbol{y} \overset{BR}{\Rightarrow} h_i(\boldsymbol{x}) \rightarrow y_i, i = 1, \ldots, m$

X_1	X_2	…	X_n	Y_1	Y_2	…	Y_m
0.12	1	…	12	0	1	…	1
2.34	9	…	-5	0	1	…	0
1.22	3	…	40	1	0	…	0
2.18	2	…	8	?	?	…	?
1.76	7	…	23	?	?	…	?

n features **m labels**

Eleftherios Spyromitros-Xioufis, PhD Student
espyromi@csd.auth.gr

Binary Relevance

- One-vs-all or Binary Relevance (BR): $h(\boldsymbol{x}) \rightarrow \boldsymbol{y} \overset{BR}{\Rightarrow} h_i(\boldsymbol{x}) \rightarrow y_i, i = 1, \ldots, m$

$\boldsymbol{h_1}$

X_1	X_2	…	X_n	Y_1	Y_2	…	Y_m
0.12	1	…	12	0	1	…	1
2.34	9	…	-5	0	1	…	0
1.22	3	…	40	1	0	…	0
2.18	2	…	8	?	?	…	?
1.76	7	…	23	?	?	…	?

n features **m labels**

Eleftherios Spyromitros-Xioufis, PhD Student
espyromi@csd.auth.gr

Binary Relevance

- One-vs-all or Binary Relevance (BR): $h(\boldsymbol{x}) \to \boldsymbol{y} \overset{BR}{\Rightarrow} h_i(\boldsymbol{x}) \to y_i, i = 1, \ldots, m$

$\boldsymbol{h_2}$

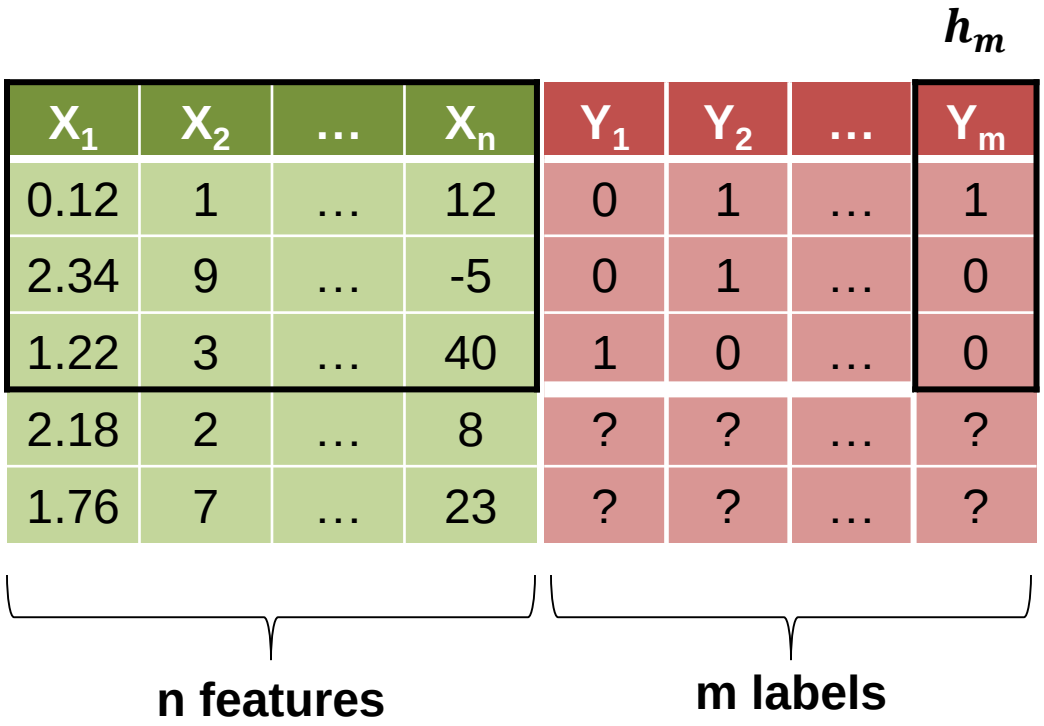
X_1	X_2	…	X_n	Y_1	Y_2	…	Y_m
0.12	1	…	12	0	1	…	1
2.34	9	…	-5	0	1	…	0
1.22	3	…	40	1	0	…	0
2.18	2	…	8	?	?	…	?
1.76	7	…	23	?	?	…	?

n features **m labels**

Eleftherios Spyromitros-Xioufis, PhD Student
espyromi@csd.auth.gr

Binary Relevance

- One-vs-all or Binary Relevance (BR): $h(\boldsymbol{x}) \to \boldsymbol{y} \overset{BR}{\Rightarrow} h_i(\boldsymbol{x}) \to y_i, i = 1, \ldots, m$



Eleftherios Spyromitros-Xioufis, PhD Student
espyromi@csd.auth.gr

Binary Relevance

- One-vs-all or Binary Relevance (BR): $h(\boldsymbol{x}) \rightarrow \boldsymbol{y} \overset{BR}{\Rightarrow} h_i(\boldsymbol{x}) \rightarrow y_i, i = 1, \ldots, m$
- Cons
  - Does not exploit label dependencies
- Pros
  - Linear scaling to the number of labels
  - Trivial to parallelize at both training and prediction
  - Flexible, can be coupled with off-the-shelf binary classifiers
  - Competitive performance
  - Good fit to the label independence assumption of the plug-in rule
    - Provides good marginals

Eleftherios Spyromitros-Xioufis, PhD Student
espyromi@csd.auth.gr

Probabilistic binary classifier

- The size of the problem is restrictive
  - Random Forests, Bagging and Boosting practically inapplicable!
- How to proceed → Large Linear Classification → LibLinear
  - Scales to millions of instances/features
  - Supports L1/L2-regularized classifiers (SVM/Logistic Regression)
- Logistic Regression was selected → probability estimates
- L2 regularizer → usually higher accuracy than L1 (poor choice ☹)
- BR + L2-regularized Logistic Regression + ML^E
  - F_1~0.740 (cost parameter tuned jointly for all labels)
  - Time ~3 days on a single core for a single train/test evaluation
  - Not much room for experimentation
- Can we do better?

Eleftherios Spyromitros-Xioufis, PhD Student
espyromi@csd.auth.gr

Feature selection using χ^2_{max}

- Data are noisy (segmentation, scanning, OCR)
  - Assumption: large part of the ~0.3M features are poor predictors
  - Try feature selection
- The χ^2_{max} criterion
  - χ^2 statistic is calculated for each feature-label combination
  - Features ranked according to max χ^2 across all labels
  - Top-k features are kept
- Feature vectors re-normalized to unit length

- Results
  - More compact: only 6% of the features!
  - More accurate: 4% better F_1 - top 10!
  - Faster: run-time reduces from 3 days to 2 hours on a single core!

n	$F_1^{internal}$
all (~0.3M)	0.740
10K	0.767
20K	**0.770**
30K	0.767

Eleftherios Spyromitros-Xioufis, PhD Student
espyromi@csd.auth.gr

Feature vector normalization

- Power normalization
  - $x = [x_1, x_2, \ldots, x_n]$ is transformed to $x^{power} = [x_1^a, x_2^a, \ldots, x_n^a]$
  - L2-normalization is re-applied on x^{power}
- Inspired from Computer Vision [4]
  - Has been applied on Bag-of-Visual-Words type histograms
    - Discounts the influence of common (visual) words
    - A variance stabilizing transform → better linear separability!
- Results
  - Small but consistent improvement for $a = 0.5$

L1-regularized Logistic Regression

n	$a = 1$	$a = 0.5$
10K	0.763	**0.767**
15K	0,766	**0.769**
20K	0.767	**0.770**
all (~0.3M)	0,778	**0.781**

[4] Jegou et al. Aggregating local image descriptors into compact codes. IEEE TPAMI 2011.

Eleftherios Spyromitros-Xioufis, PhD Student
espyromi@csd.auth.gr

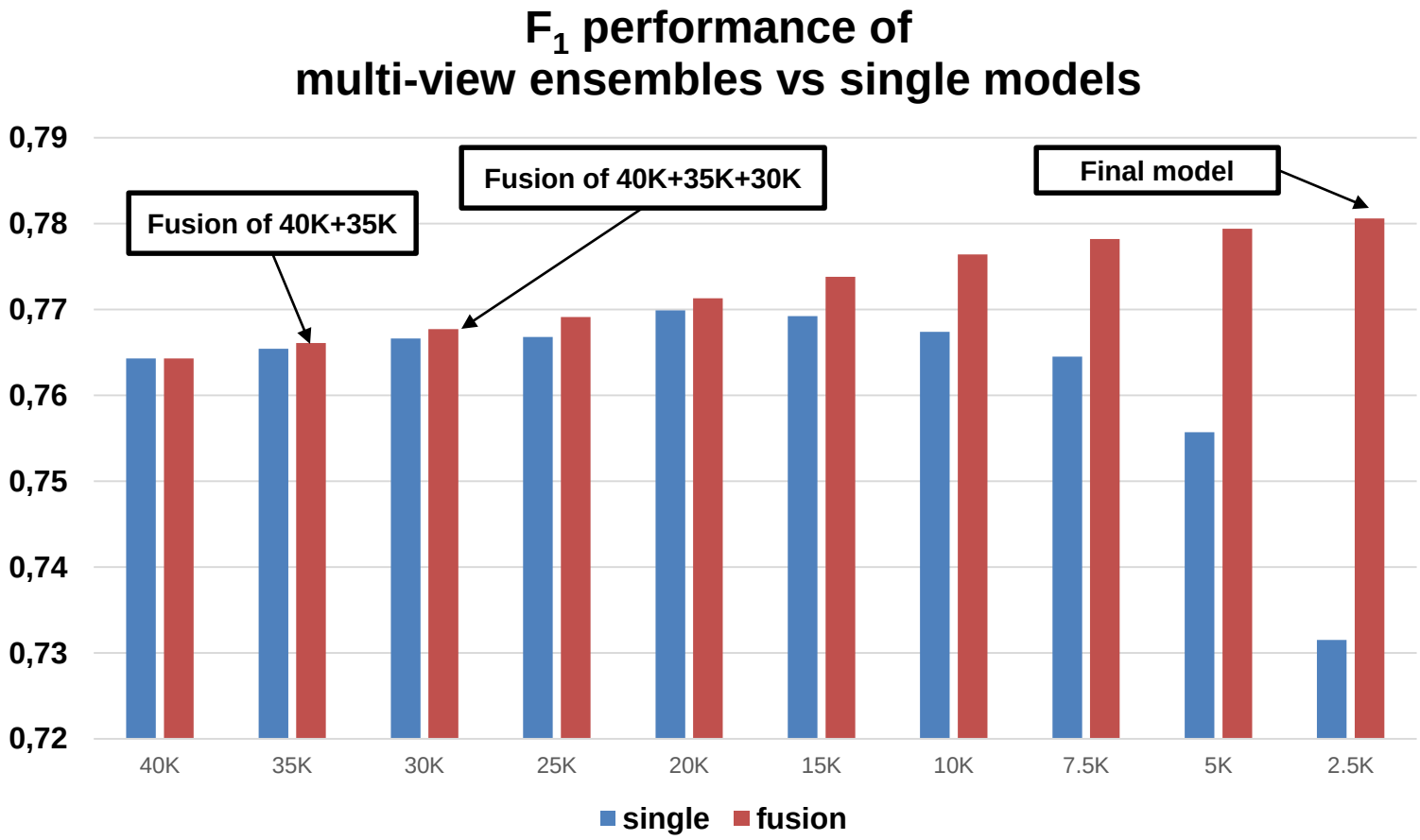
Final model: a multi-view ensemble

- Motivation
  - Approximately same performance with different top-k χ^2_{max} features!
  - What if we combine these models?
- Combination by simple averaging of predictions:
  - $h^{mv}(x) = \frac{1}{N}(h(x^{topk_1}) + \cdots + h(x^{topk_N}))$
  - Final predictions obtained by ML^E
- top-k common words also used in [5]
  - Interpretation: different subsets work better on different labels

[5] K. Sechidis. Multi-label machine learning algorithms for automated image annotation. MSc Thesis.

Eleftherios Spyromitros-Xioufis, PhD Student
espyromi@csd.auth.gr

Final model: a multi-view ensemble



Eleftherios Spyromitros-Xioufis, PhD Student
espyromi@csd.auth.gr

What didn't work..

- Taking example order into account
  - 1st approach: train using last x% of the training examples
  - 2nd approach: train using all examples but assign larger weight to the latest
- Could work under 2 conditions:
  - Documents chronologically ordered within training set
  - Concept drift:
    - Distribution of latest training examples closer to test distribution
  - Preliminary experiments showed performance deterioration

Eleftherios Spyromitros-Xioufis, PhD Student
espyromi@csd.auth.gr

Post-contest experiments

- Preferring L2 over L1 regularization proved to be wrong!
- Better performance → “S 6” instead of “S 0” as LibLinear solver ☺
  - Observations:
    - using all features is now better than χ^2_{max} feature selection
    - power normalization and multi-view ensemble help as before
    - performance increases from ~ 0.78 to ~ 0.79
- How to improve
  - More sophisticated multi-label algorithms
  - Better binary classifiers e.g. linear SVMs with probability outputs
  - Blending different algorithms
  - Better exploitation of distribution changes

Eleftherios Spyromitros-Xioufis, PhD Student
espyromi@csd.auth.gr

Conclusions

- Conceptually simple, efficient and effective solution
- Everything programmed in Java → easy to deploy
- Intuitive decisions were theoretically justified

Kaggle master badge ☺

Eleftherios Spyromitros-Xioufis, PhD Student
espyromi@csd.auth.gr

THANK YOU!
QUESTIONS?

If you are interested contact me at: espyromi@csd.auth.gr

Eleftherios Spyromitros-Xioufis, PhD Student
espyromi@csd.auth.gr