Northeastern University

# A Pipeline for Optimizing F1-Measure in Multi-Label Text Classification

Bingyu Wang, Cheng Li, Virgil Pavlu, Javed Aslam {rainicy, chengli, vip, jaa}@ ccs.neu.edu

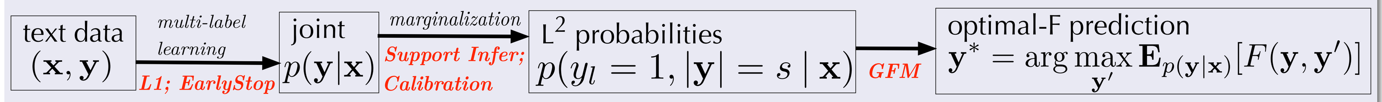
IEEE ICMLA 2018

## Problem

**Multi-label classification**: assign each instance with multiple labels, e.g. a news is related to *Politics* and *Election*.
**Challenge**: how to incorporate label dependencies in an efficient way in order to improve F1-performance:

$$F(\mathbf{y}, \mathbf{y}') = \frac{2\sum_{l=1}^{L} y_l y_l'}{\sum_{l=1}^{L} y_l + \sum_{l=1}^{L} y_l'}$$

## Proposed Pipeline




The proposed pipeline takes any probabilistic multi-label classifiers in general and improves their F1-measure with careful training regularization and a new prediction strategy:

- training (**L**): L1 with L2 regularization, also called ElasticNet, is essential for high dimensional documents classification problem: L2 spreads weights to correlated features, and L1 shrinks some irrelevant features to zeros.
- prediction:
  - General F-Measure Maximizer (**G**): an algorithm to optimize F1-measure with marginal distributions of the form:
    $$p(y_l = 1, |\mathbf{y}| = s \mid \mathbf{x}), \quad \forall l, s \in \{1, ..., L\}$$
  - Support Inference (**S**): consider those label combinations only appearing in training set and marginalizes over their probabilities.
  - Calibration (**C**): calibrate the probabilities estimated from the support inference.

## Applied Approaches

**Binary Relevance (BR)**, predicts each binary label independently: $p(\mathbf{y}|\mathbf{x}) = \prod_{l=1}^{L} p(y_l|\mathbf{x})$
**Probabilistic Classifier Chain (PCC)** constructs a chain of binary classifiers for labels:

$$p(\mathbf{y}|\mathbf{x}) = p(y_1|\mathbf{x})p(y_2|\mathbf{x}, y_1)\cdots p(y_L|\mathbf{x}, y_1, .., y_{L-1})$$

**Pair-wise CRF** specifies label dependency with CRF:

$$p(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})}\exp\{\sum_{l=1}^{L}\sum_{d=1}^{D} w_{ld}x_d \mathbb{1}[y_l = 1]$$
$$+\sum_{l=1}^{L}\sum_{m=1}^{L}(w_{lm1}\mathbb{1}[y_l = 0, y_m = 0] + w_{lm2}\mathbb{1}[y_l = 0, y_m = 1]$$
$$+w_{lm3}\mathbb{1}[y_l = 1, y_m = 0] + w_{lm4}\mathbb{1}[y_l = 1, y_m = 1])\}$$

**CBM** estimates a joint probability by a mixture of conditional Bernoulli:

$$p(\mathbf{y}|\mathbf{x}) = \sum_{k=1}^{K}\pi(z = k|\mathbf{x})\prod_{l=1}^{L} b(y_l|\mathbf{x}, z = k)$$

**GFM** maximizes the expected F1-measure during the prediction:

$$\mathbf{y}^* = \arg\max_{\mathbf{y}'}\sum_{\mathbf{y}} p(\mathbf{y}|\mathbf{x}) \cdot F(\mathbf{y}, \mathbf{y}')$$

## Datasets Characteristics

| | BIBTEX | IMDB | OHSUMED | RCV1 | WISE | WIPO |
|---|---|---|---|---|---|---|
| **domain** | bkmark | genre | medical | news | articles | patent |
| **source** | Mulan | crawled* | MEKA* | Mulan | WISE2014 | HRSVM |
| **labels** | 159 | 27 | 23 | 101 | 203 | 188 |
| **label sets** | 2,058 | 2,122 | 1,042 | 494 | 3,536 | 155 |
| **features** | 1,836 | 27,228 | 16,344 | 47,236 | 301,561 | 74,435 |
| **instances** | 7,395 | 34,157 | 13,929 | 6,000 | 64,857 | 1,710 |
| **cardinality** | 2.40 | 2.52 | 1.66 | 3.23 | 1.45 | 4.00 |
| **inst/label** | 112 | 2537 | 1007 | 188 | 463 | 36 |

Note: cardinality = average number of labels per instance; inst/label = the average number of training instances per label.

## Main Analysis with the proposed pipeline

Table: F-measure on test w/ and w/o L1(L), Support Inference(S), GFM(G) and Calibration(C)

| Data | Model | Standard | SG | L | LS | LG | LSG | LSCG |
|---|---|---|---|---|---|---|---|---|
| BIBT | BR | 37.8 | 44.5 | 39.8 | 44.4 | 40.2 | 45.4 | **48.1** |
| | CRF\ L1 | - | - | - | 46.5 | - | 49.4 | **49.5** |
| | PCC | 37.4 | 45.3 | 39.5 | 45.0 | 40.1 | 47.3 | **48.2** |
| | CBM | 44.0 | 45.9 | 45.3 | 46.9 | 40.4 | 49.5 | **50.4*** |
| IMDB | BR | 59.4 | 61.8 | 59.6 | 59.7 | 61.0 | 61.4 | **63.8** |
| | CRF\ L1 | - | - | - | 63.0 | - | 66.6 | **67.1*** |
| | PCC | 59.6 | 63.9 | 60.1 | 60.2 | 61.5 | 62.8 | **64.4** |
| | CBM | 61.6 | 65.1 | 62.2 | 62.2 | 64.8 | 65.2 | **66.2** |
| OHSU | BR | 60.2 | 67.9 | 63.6 | 68.0 | 64.3 | 69.1 | **71.0** |
| | CRF\ L1 | - | - | - | 66.4 | - | 69.6 | **70.5** |
| | PCC | 62.5 | 70.1 | 64.7 | 68.4 | 65.8 | 70.4 | **72.1** |
| | CBM | 68.7 | 70.3 | 69.5 | 70.3 | 65.4 | 71.7 | **72.6*** |
| RCV1 | BR | 72.1 | 73.7 | 73.8 | 74.6 | 74.9 | 75.1 | **76.1** |
| | CRF\ L1 | - | - | - | 74.4 | - | 75.8 | **76.1** |
| | PCC | 71.0 | 73.6 | 72.7 | 72.8 | 74.3 | 74.1 | **74.4** |
| | CBM | 76.6 | 77.3 | 77.3 | 78.5 | 77.9 | **79.2*** | 78.7 |
| WISE | BR | 68.0 | 77.3 | 72.8 | 79.0 | 73.0 | 79.3 | **80.1** |
| | CRF\ L1 | - | - | - | 77.7 | - | 79.0 | **79.4** |
| | PCC | 70.7 | 76.0 | 74.6 | 76.7 | 77.1 | **78.0** | - |
| | CBM | 77.9 | 78.6 | 79.8 | 79.8 | 73.6 | 80.3 | **81.5*** |
| WIPO | BR | 63.4 | 71.2 | 69.5 | 73.2 | 70.0 | **74.0** | 68.0 |
| | CRF\ L1 | - | - | - | 70.3 | - | 72.2 | **72.5** |
| | PCC | 68.8 | 71.5 | 70.2 | 70.4 | 70.6 | **72.3** | 54.6 |
| | CBM | 63.0 | 70.8 | 69.6 | 72.5 | 70.3 | **74.3*** | 71.3 |

Note:**bold**: best in row; *****: best in dataset; "-": N/A.

## Model size and feature used with L1&L2

| Data | BR | | | CBM | | |
|---|---|---|---|---|---|---|
| | L2 Only model(MB) | L1L2 feature used | L1L2 model used | L2 only model | L1L2 feature used | L1L2 model used |
| BIBT | 7 | 100% | 26% | 135 | 100% | 4% |
| IMDB | 20 | 66% | 21% | 355 | 99% | 10% |
| OHSU | 10 | 53% | 34% | 177 | 68% | 6% |
| RCV1 | 48 | 70% | 12% | 910 | 77% | 2% |
| WISE | 1.4(G) | 14% | 1% | 13(G) | 24% | <1% |
| WIPO | 294 | 42% | 2% | 6G | 77% | 2% |

Note: Percentages of the L2 model/feature size after adding L1.

## CRF w/ and w/o label-label pair.

| | BIBT | | IMDB | | OHSU | | RCV1 | | WISE | | WIPO | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| pairwise | w/o | w/ | w/o | w/ | w/o | w/ | w/o | w/ | w/o | w/ | w/o | w/ |
| CRF w/o GFM | 46.9 | 46.5 | 61.3 | 63.0 | 66.1 | 66.4 | 73.8 | 74.4 | 78.2 | 77.7 | 70.7 | 70.3 |
| CRF w/ GFM | 49.4 | 49.4 | 66.1 | 66.6 | 69.8 | 69.6 | 75.8 | 75.8 | 79.4 | 79.0 | 71.8 | 72.2 |

## F-measure comparisons with other methods

| Method | BIBT | IMDB | OHSU | RCV1 | WISE | WIPO |
|---|---|---|---|---|---|---|
| BR SVM + L2 | 37.8 | 59.9 | 60.9 | 73.4 | 70.0 | 64.7 |
| BR SVM + L1 | 39.3 | 59.0 | 63.5 | 73.0 | 70.0 | 68.1 |
| BR LR + L2 | 38.1 | 60.0 | 61.1 | 72.3 | 68.6 | 64.3 |
| BR LR + L1 | 39.0 | 60.5 | 61.4 | 73.4 | 70.4 | 68.7 |
| LIFT | 31.5 | - | 54.4 | 70.2 | - | 61.6 |
| SPEN + L2 | 39.0 | 61.1 | 61.7 | 65.3 | - | 65.9 |
| PDsparse+L1L2 | 40.7 | 62.3 | 67.3 | 75.0 | 74.5 | 67.5 |
| CFT | 23.5 | - | - | 53.5 | - | 62.7 |
| CLEMS | 42.5 | - | 52.6 | 72.4 | - | 67.1 |
| LSF | 43.9 | 59.8 | 65.0 | 73.6 | 76.7 | 71.1 |
| BR+LSCG† | 48.1 | 63.8 | 71.0 | 76.1 | 80.1 | 68.0 |
| CRF+LSCG† | 49.5 | **67.1** | 70.5 | 76.1 | 79.4 | **72.5** |
| CBM+LSCG† | **50.4** | 66.2 | **72.6** | **78.7** | **81.5** | 71.3 |

Note: †: our method; '-': indicates failed runs with 56 core and 256GB RAM.