On the Stratification of Multi-Label Data

Konstantinos Sechidis, **Grigorios Tsoumakas**, Ioannis Vlahavas

Machine Learning & Knowledge Discovery Group
Department of Informatics
Aristotle University of Thessaloniki
Greece

Stratified Sampling

- Sampling plays a key role in practical machine learning and data mining
  - Exploration and efficient processing of vast data
  - Generation of training, validation and test sets for *accuracy estimation*, *model selection*, *hyper-parameter selection* and *overfitting avoidance* (e.g. reduced error pruning)
- The stratified version of sampling is typically used in classification tasks
  - The proportion of the examples of each class in a sample of a dataset follows that of the full dataset
  - It has been found to improve standard cross-validation both in terms of bias and variance of estimate (Kohavi, 1995)

Stratifying Multi-Label Data

- Instances associated with a subset of a fixed set of labels

Male, *Horse*, *Natural*, *Animals*, *Sunny*, *Day*, *Mountains*, *Clouds*, *Sky*, *Plants*, *Outdoor*

Stratifying Multi-Label Data

- Random sampling is typically used in the literature
- We consider two main approaches for the stratification of multi-label data
  - Stratified sampling based on labelsets (label combinations)
    - The number of labelsets is often quite large and each labelset is associated with very few examples, rendering this approach impractical
  - Set as goal the maintenance of the distribution of positive and negative examples of each label
    - This views the problem independently for each label
    - It cannot be achieved by simple independent stratification of each label, as the produced subsets need to be the same
    - Our solution: iterative stratification of labels

Stratification Based on Labelsets

instance	λ_1	λ_2	λ_3
i_1	1	0	1
i_2	0	0	1
i_3	0	1	0
i_4	1	0	0
i_5	0	1	1
i_6	1	1	0
i_7	1	0	1
i_8	1	0	1
i_9	0	0	1

labelset
5
<u>1</u>
2
4
3
6
5
5
<u>1</u>

1st Fold

2nd Fold

3rd Fold

Stratification Based on Labelsets

instance	λ_1	λ_2	λ_3
i_1	1	0	1
i_2	0	0	1
i_3	0	1	0
i_4	1	0	0
i_5	0	1	1
i_6	1	1	0
i_7	1	0	1
i_8	1	0	1
i_9	0	0	1

labelset
5
1
2
4
3
6
5
5
1

1st Fold				
i_1	1	0	1	**5**
i_2	0	0	1	1
i_3	0	1	0	2

2nd Fold				
i_7	1	0	1	**5**
i_9	0	0	1	1
i_4	1	0	0	4

3rd Fold				
i_8	1	0	1	**5**
i_5	0	1	1	3
i_6	1	1	0	6

Statistics of Multi-Label Data

dataset	labels	examples	labelsets	labelsets / examples	examples per labelset			examples per label		
					min	avg	max	min	avg	max
Scene	6	2407	15	**0.01**	1	**160**	405	364	431	533
Emotions	6	593	27	**0.05**	1	**22**	81	148	185	264
TMC2007	22	28596	1341	**0.05**	1	**21**	2486	441	2805	16173
Genbase	27	662	32	**0.05**	1	**21**	170	1	31	171
Yeast	14	2417	198	**0.08**	1	**12**	237	34	731	1816
Medical	45	978	94	**0.1**	1	**10**	155	1	27	266
Mediamill	101	43907	6555	**0.15**	1	**7**	2363	31	1902	33869
Bookmarks	208	87856	18716	**0.21**	1	**5**	6087	300	857	6772
Bibtex	159	7395	2856	**0.39**	1	**3**	471	51	112	1042
Enron	53	1702	753	**0.44**	1	**2**	163	1	108	913
Corel5k	374	5000	3175	**0.64**	1	**2**	55	1	47	1120
ImageCLEF2010	93	8000	7366	**0.92**	1	**1**	32	12	1038	7484
Delicious	983	16105	15806	**0.98**	1	**1**	19	21	312	6495

Iterative Stratification Algorithm

- Select the label with the fewest remaining examples
  - If rare labels are not examined in priority, they may be distributed in an undesired way, beyond subsequent repair
  - For frequent labels, we have the chance to modify the current distribution towards the desired one in a subsequent iteration, due to the availability of more examples
- For each example of this label, select the subset with
  - The largest desired number of examples for this label
  - The largest desired number of examples, in case of ties
  - Further ties are broken randomly
- Update statistics
  - Desired number of examples per label at each subset

Note: No hard constrain on the desired number of examples

Example

Instance	λ_1	λ_2	λ_3
i_1	1	0	1
i_2	0	0	1
i_3	0	1	0
i_4	1	0	0
i_5	0	1	1
i_6	1	1	0
i_7	1	0	1
i_8	1	0	1
i_9	0	0	1
sum	**5**	**3**	**6**

1st Fold			
desired	**1.7**	**1**	**2**

2nd Fold			
desired	**1.7**	**1**	**2**

3rd Fold			
desired	**1.7**	**1**	**2**

Example

Instance	λ_1	λ_2	λ_3
i_1	1	0	1
i_2	0	0	1
i_3	0	1	0
i_4	1	0	0
i_5	0	1	1
i_6	1	1	0
i_7	1	0	1
i_8	1	0	1
i_9	0	0	1
sum	**5**	**3**	**6**

<u>Firstly</u>
Distribute the positive examples of λ_2

1st Fold			
desired	**1.7**	**1**	**2**

2nd Fold			
desired	**1.7**	**1**	**2**

3rd Fold			
desired	**1.7**	**1**	**2**

Example

<u>Firstly</u>
Distribute the positive examples of λ_2

Instance	λ_1	λ_2	λ_3
i_1	1	0	1
i_2	0	0	1
i_4	1	0	0
i_5	0	1	1
i_6	1	1	0
i_7	1	0	1
i_8	1	0	1
i_9	0	0	1
sum	**5**	**2**	**6**

1st Fold			
i_3	0	1	0
desired	**1.7**	**0**	**2**

2nd Fold			
desired	**1.7**	**1**	**2**

3rd Fold			
desired	**1.7**	**1**	**2**

Example

Instance	λ_1	λ_2	λ_3
i_1	1	0	1
i_2	0	0	1
i_4	1	0	0
i_6	1	1	0
i_7	1	0	1
i_8	1	0	1
i_9	0	0	1
sum	**5**	**1**	**5**

<u>Firstly</u>
Distribute the positive examples of λ_2

1st Fold			
i_3	0	1	0
desired	**1.7**	**0**	**2**

2nd Fold			
desired	**1.7**	**1**	**2**

3rd Fold			
i_5	0	1	1
desired	**1.7**	**0**	**1**

Example

<u>Firstly</u>
Distribute the positive examples of λ_2

Instance	λ_1	λ_2	λ_3
i_1	1	0	1
i_2	0	0	1
i_4	1	0	0
i_7	1	0	1
i_8	1	0	1
i_9	0	0	1
sum	**4**	**-**	**5**

1st Fold

i_3	0	1	0
desired	**1.7**	**0**	**2**

2nd Fold

i_6	1	1	0
desired	**0.7**	**0**	**2**

3rd Fold

i_5	0	1	1
desired	**1.7**	**0**	**1**

Example

<u>Secondly</u>

Distribute the positive examples of λ_1

Instance	λ_1	λ_2	λ_3
i_1	1	0	1
i_2	0	0	1
i_4	1	0	0
i_7	1	0	1
i_8	1	0	1
i_9	0	0	1
sum	**4**	**-**	**5**

1st Fold			
i_3	0	1	0
desired	**1.7**	**0**	**2**

2nd Fold			
i_6	1	1	0
desired	**0.7**	**0**	**2**

3rd Fold			
i_5	0	1	1
desired	**1.7**	**0**	**1**

Example

<u>Secondly</u>

Distribute the positive examples of λ_1

Instance	λ_1	λ_2	λ_3
i_2	0	0	1
i_4	1	0	0
i_7	1	0	1
i_8	1	0	1
i_9	0	0	1
sum	**3**	**-**	**4**

1st Fold			
i_3	0	1	0
i_1	1	0	1
desired	**0.7**	**0**	**1**

2nd Fold			
i_6	1	1	0
desired	**0.7**	**0**	**2**

3rd Fold			
i_5	0	1	1
desired	**1.7**	**0**	**1**

Example

<u>Secondly</u>

Distribute the positive examples of λ_1

Instance	λ_1	λ_2	λ_3
i_2	0	0	1
i_7	1	0	1
i_8	1	0	1
i_9	0	0	1
sum	**2**	**-**	**4**

1st Fold			
i_3	0	1	0
i_1	1	0	1
desired	**0.7**	**0**	**1**

2nd Fold			
i_6	1	1	0
desired	**0.7**	**0**	**2**

3rd Fold			
i_5	0	1	1
i_4	1	0	0
desired	**0.7**	**0**	**1**

Example

<u>Secondly</u>

Distribute the positive examples of λ_1

Instance	λ_1	λ_2	λ_3
i_2	0	0	1
i_8	1	0	1
i_9	0	0	1
sum	**1**	**-**	**3**

1st Fold			
i_3	0	1	0
i_1	1	0	1
desired	**0.7**	**0**	**1**

2nd Fold			
i_6	1	1	0
i_7	1	0	1
desired	**-0.3**	**0**	**1**

3rd Fold			
i_5	0	1	1
i_4	1	0	0
desired	**0.7**	**0**	**1**

Example

<u>Secondly</u>

Distribute the positive examples of λ_1

Instance	λ_1	λ_2	λ_3
i_2	0	0	1
i_9	0	0	1
sum	**-**	**-**	**2**

1st Fold			
i_3	0	1	0
i_1	1	0	1
i_8	1	0	1
desired	**-0.3**	**0**	**0**

2nd Fold			
i_6	1	1	0
i_7	1	0	1
desired	**-0.3**	**0**	**1**

3rd Fold			
i_5	0	1	1
i_4	1	0	0
desired	**0.7**	**0**	**1**

Example

<u>Thirdly</u>

Distribute the positive examples of λ_3

Instance	λ_1	λ_2	λ_3
i_2	0	0	1
i_9	0	0	1
sum	**-**	**-**	**2**

1st Fold			
i_3	0	1	0
i_1	1	0	1
i_8	1	0	1
desired	**-0.3**	**0**	**0**

2nd Fold			
i_6	1	1	0
i_7	1	0	1
desired	**-0.3**	**0**	**1**

3rd Fold			
i_5	0	1	1
i_4	1	0	0
desired	**0.7**	**0**	**1**

Example

Thirdly

Distribute the positive examples of λ_3

Instance	λ_1	λ_2	λ_3
i_9	0	0	1
sum	**-**	**-**	**1**

1^{st} Fold			
i_3	0	1	0
i_1	1	0	1
i_8	1	0	1
desired	**-0.3**	**0**	**0**

2^{nd} Fold			
i_6	1	1	0
i_7	1	0	1
i_2	0	0	1
desired	**-0.3**	**0**	**0**

3^{rd} Fold			
i_5	0	1	1
i_4	1	0	0
desired	**0.7**	**0**	**1**

Example

<u>Thirdly</u>

Distribute the positive examples of λ_3

Instance	λ_1	λ_2	λ_3
sum	**-**	**-**	**-**

1st Fold			
i_3	0	1	0
i_1	1	0	1
i_8	1	0	1
desired	**-0.3**	**0**	**0**

2nd Fold			
i_6	1	1	0
i_7	1	0	1
i_2	0	0	1
desired	**-0.3**	**0**	**0**

3rd Fold			
i_5	0	1	1
i_4	1	0	0
i_9	0	0	1
desired	**0.7**	**0**	**0**

The Triggering Event

- Implementation of evaluation software
  - Stratification of multi-label data concerned us a while ago during the development of the Mulan open-source library
- However, a more practical issue triggered this work
  - During our participation at ImageCLEF 2010, x-validation experiments led to subsets without positive examples for some labels, and problems in the calculation of the main evaluation measure of the challenge, Mean Avg Precision

Subsets Without Label Examples

- When can this happen?
  - When there are rare labels
- Problems in calculation of evaluation measures
  - A test set without positive examples for a label (fn=tp=0) renders *recall* undefined, and so gets F_1, *AUC* and *MAP*
  - Furthermore, if the model is correct (fp=0) then *precision* is undefined

		Predicted	
		negative	positive
Actual	negative	*tn*	*fp*
	positive	*fn*	*tp*

Recall: tp/(tp+fn)
Precision: tp/(tp+fp)

Comparison of the Approaches

intends to maintain ***joint*** distribution

random

1st Fold				
i_1	1	0	1	5
i_2	0	0	1	1
i_3	0	1	0	2
2nd Fold				
i_4	1	0	0	4
i_5	0	1	1	3
i_6	1	1	0	6
3rd Fold				
i_7	1	**0**	1	**5**
i_8	1	**0**	1	**5**
i_9	0	**0**	1	1

based on labelsets

1st Fold				
i_1	1	0	1	5
i_2	0	0	1	1
i_3	0	1	0	2
2nd Fold				
i_7	1	**0**	1	5
i_9	0	**0**	1	1
i_4	1	**0**	0	4
3rd Fold				
i_8	1	0	1	5
i_5	0	1	1	3
i_6	1	1	0	6

iterative

1st Fold				
i_3	0	1	0	2
i_1	1	0	1	**5**
i_8	1	0	1	**5**
2nd Fold				
i_6	1	1	0	6
i_7	1	0	1	5
i_2	0	0	1	1
3rd Fold				
i_5	0	1	1	3
i_4	1	0	0	4
i_9	0	0	1	1

intends to maintain ***marginal*** distribution

Experiments

- Sampling approaches
  - Random (**R**)
  - Stratified sampling based on labelsets (**L**)
  - Iterative stratification algorithm (**I**)
- We experiment on 13 multi-label datasets
  - 10-fold CV on datasets with up to 15k examples and
  - Holdout (2/3 for training and 1/3 for testing) on larger ones
- Experiments are repeated 5 times with different random orderings of the training examples
  - Presented results are averages over these 5 experiments

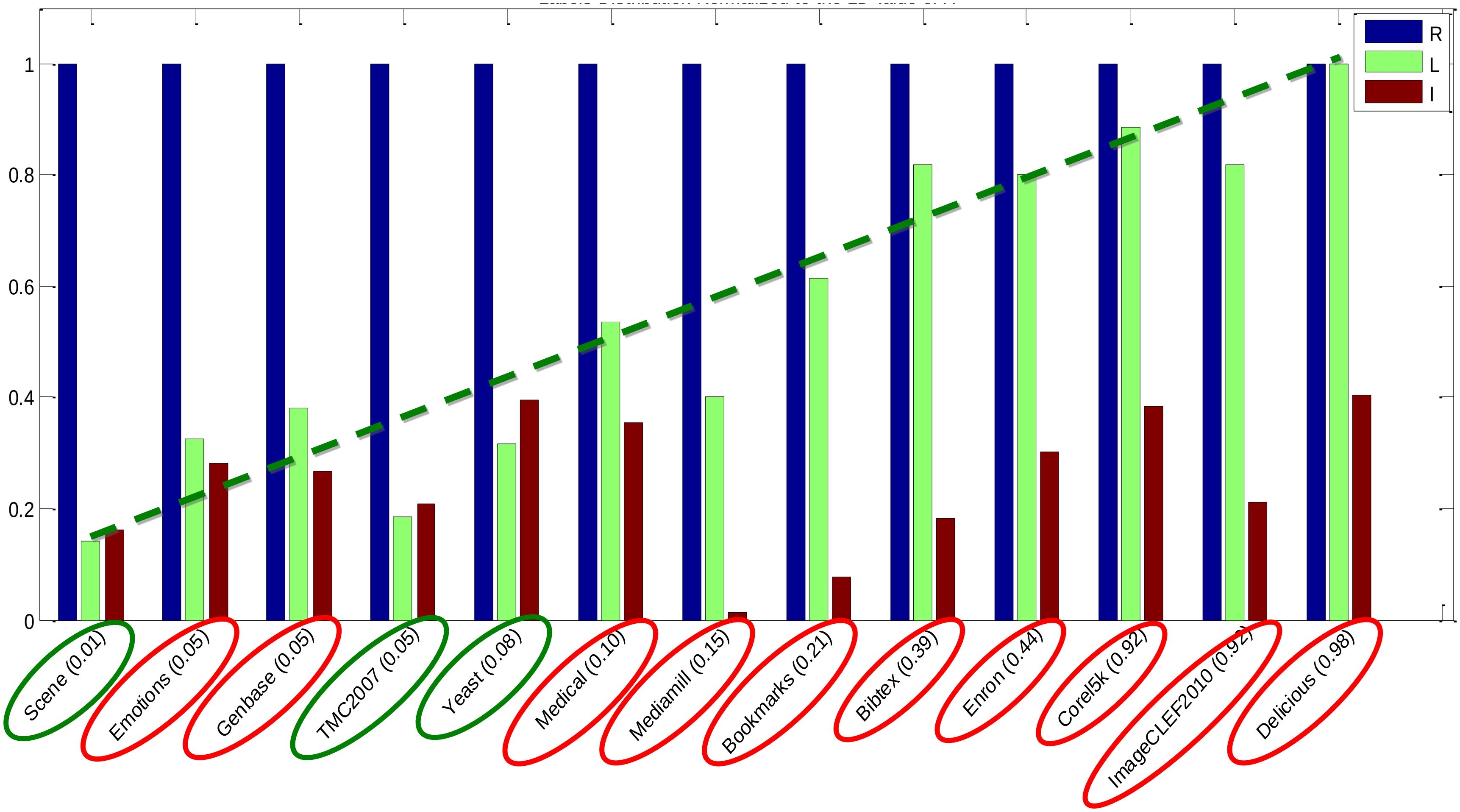
Distribution of Labels & Examples

- Notation
  - q labels, k subsets, c_j desired examples in subset j,
  - D^i: set of examples of label i, S_j: set of examples in subset j
  - S^i_j: set of examples of label i in subset j
- Labels distribution (LD) and examples distribution (ED)

$$LD = \frac{1}{q}\sum_{i=1}^{q}\left(\frac{1}{k}\sum_{j=1}^{k}\left|\frac{\left|S_j^i\right|}{\left|S_j\right|-\left|S_j^i\right|}-\frac{\left|D^i\right|}{\left|D\right|-\left|D^i\right|}\right|\right) \qquad ED = \frac{1}{k}\sum_{j=1}^{k}\left|\left|S_j\right|-c_j\right|$$

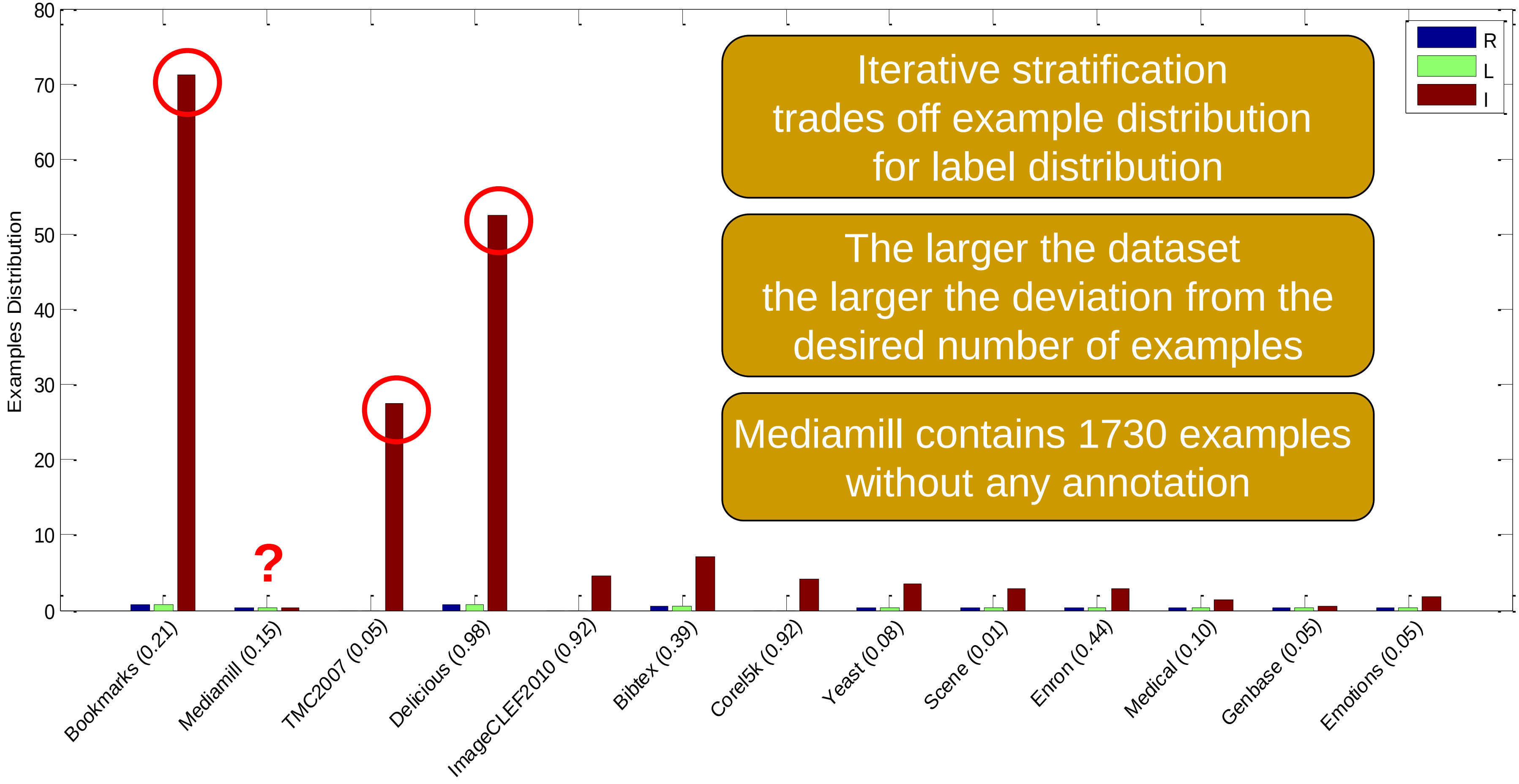
- Subsets without positive examples
  - Number of folds that contain at least one label with zero positive examples (*FZ*), number of fold-label pairs with zero positive examples (*FLZ*)

Labels Distribution (normalized)



Datasets are sorted in increasing order of #labelsets/#examples

Examples Distribution



Datasets are sorted in decreasing order of #examples

Subsets Without Label Examples

dataset	labels	labelsets / examples	examples per label			FZ			FLZ		
			min	avg	max	R	L	I	R	L	I
Scene	6	**0.01**	364	431	533	0	0	0	0	0	0
Emotions	6	**0.05**	148	185	264	0	0	0	0	0	0
Genbase	27	**0.05**	1	31	171	10	10	10	90	77	74
Yeast	14	**0.08**	34	731	1816	1	0	0	1	0	0
Medical	45	**0.1**	1	27	266	10	10	10	203	179	173
Bibtex	159	**0.39**	51	112	1042	1	1	0	1	1	0
Enron	53	**0.44**	1	108	913	10	10	10	95	88	47
Corel5k	374	**0.64**	1	47	1120	10	10	10	1140	1118	788
ImageCLEF2010	93	**0.92**	12	1038	7484	4	4	0	4	0	0

- Iterative stratification produces the lowest FZ & FLZ in all datasets

- All schemes fail in Genbase, Medical, Enron and Corel5k due to label rarity

- All schemes do well in Scene, Emotions, where examples per label abound

- Only iterative stratification does well in Bibtex and ImageCLEF2010

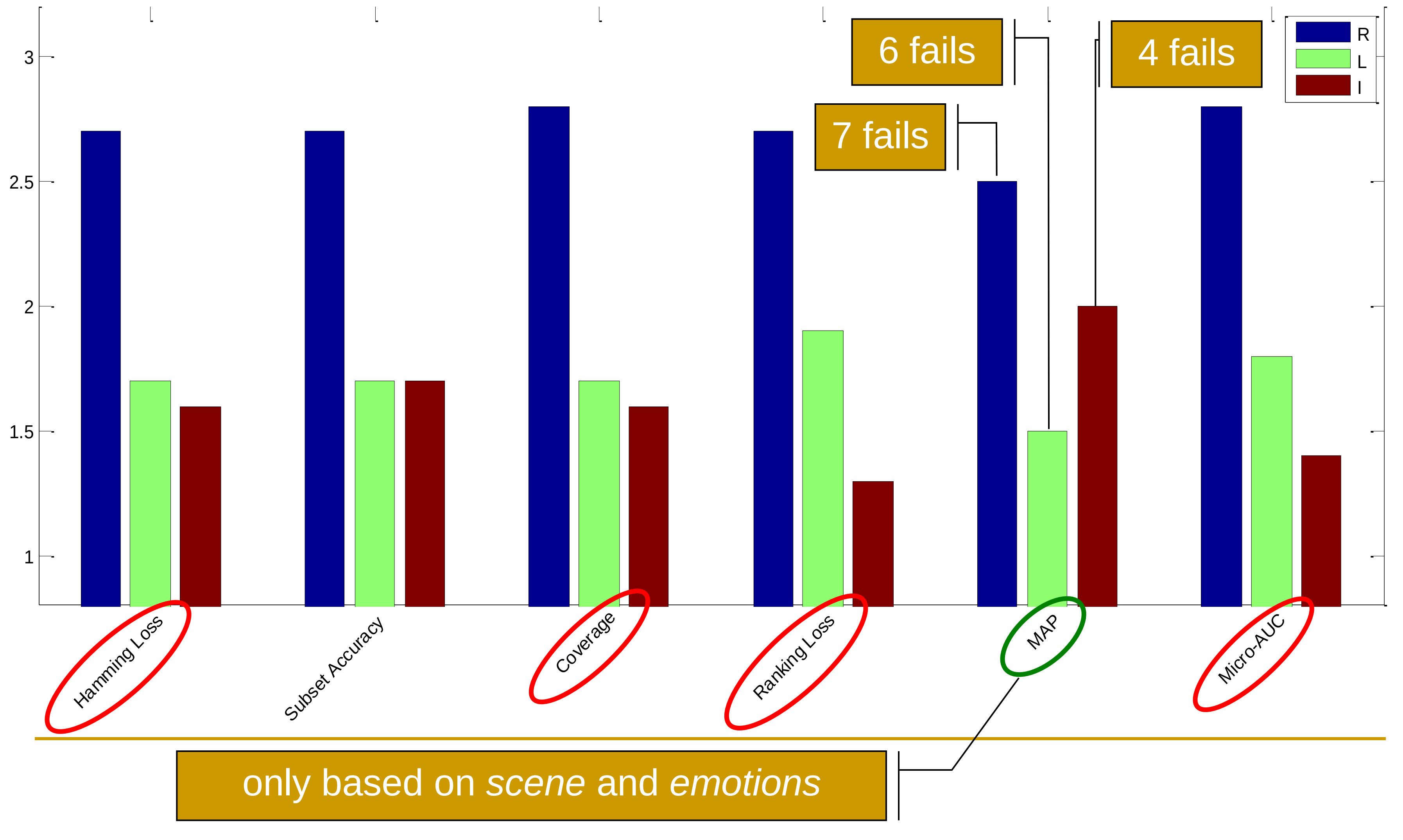
Variance of 10-fold CV Estimates

- Algorithms
  - Binary Relevance (one-versus-rest)
  - Calibrated Label Ranking (Fürnkranz et al., 2008)
    - Combination of pairwise and one-versus-rest models
    - Considers label dependencies
- Measures

Measure	Required type of output
Hamming Loss	Bipartition
Subset Accuracy	Bipartition
Coverage	Ranking
Ranking Loss	Ranking
Mean Average Precision	Probabilities
Micro-averaged AUC	Probabilities

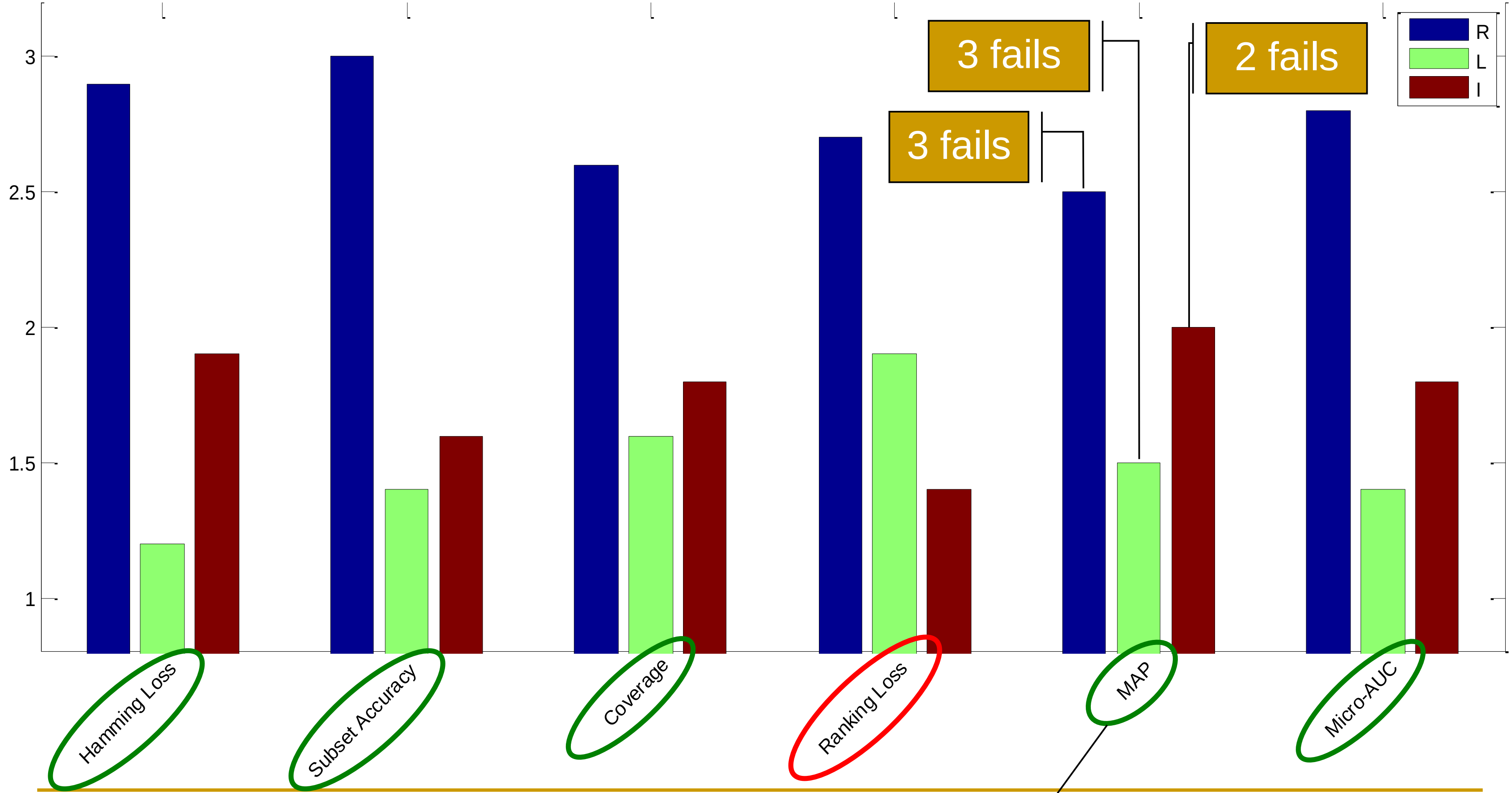
Average Ranking for BR (1/3)

- On all 9 datasets



Average Ranking for BR (2/3)

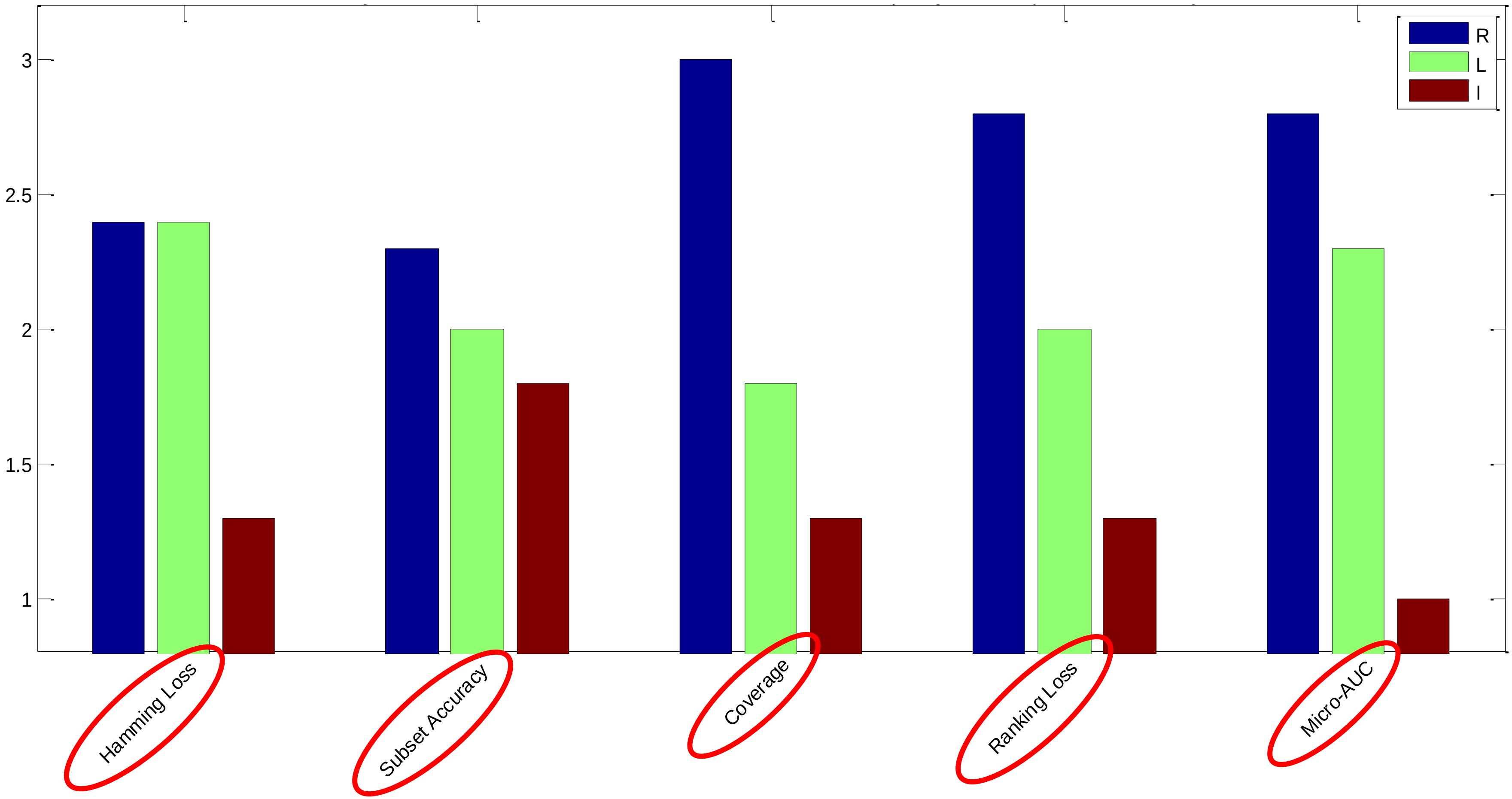
- On 5 datasets where #labelsets/#examples ≤ 0.1



only based on *scene* and *emotions*

Average Ranking for BR (3/3)

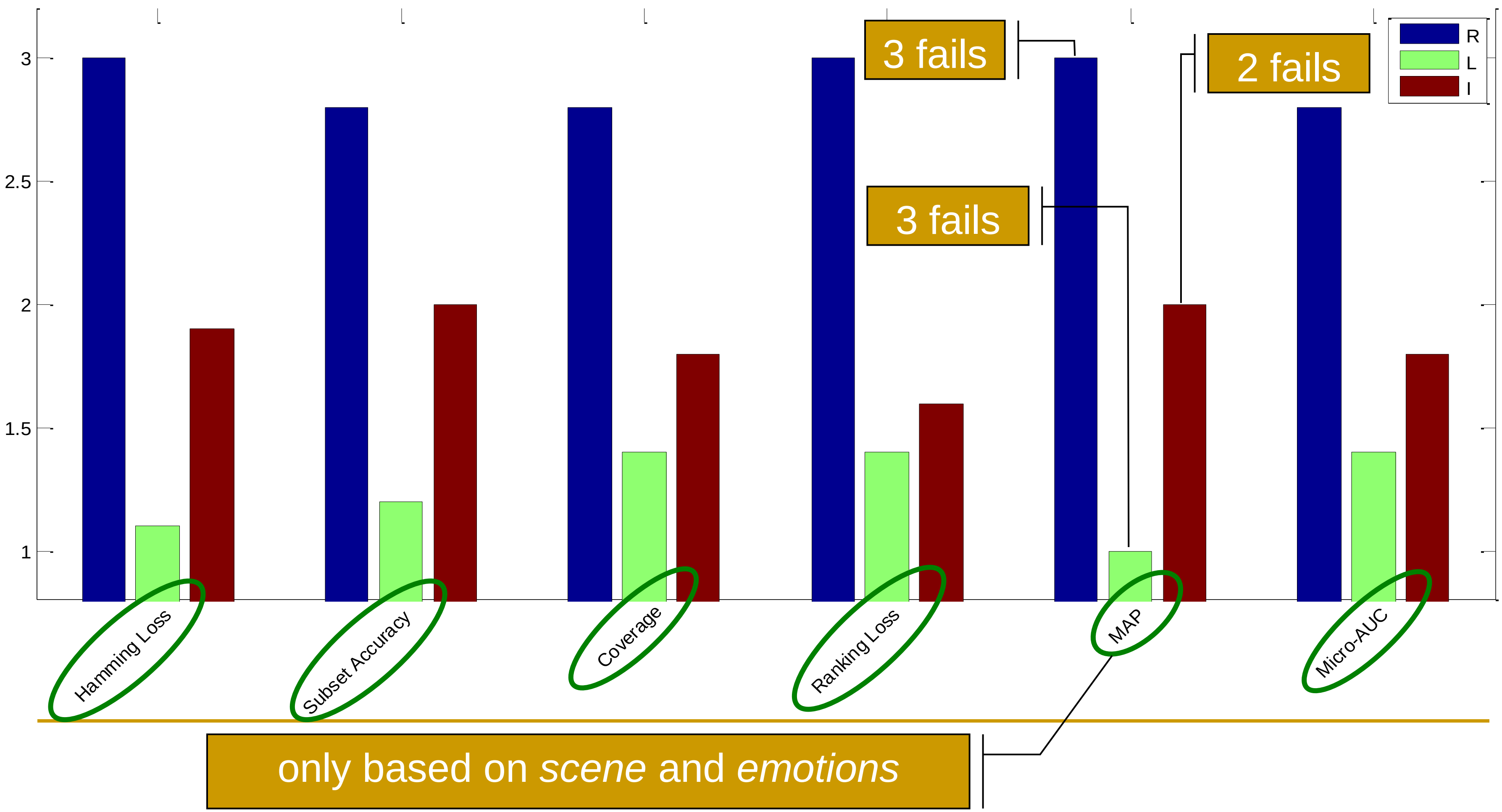
- On 4 datasets where #labelsets/#examples ≥ 0.39



Fails in MAP – R: 4, L: 4, I: 2

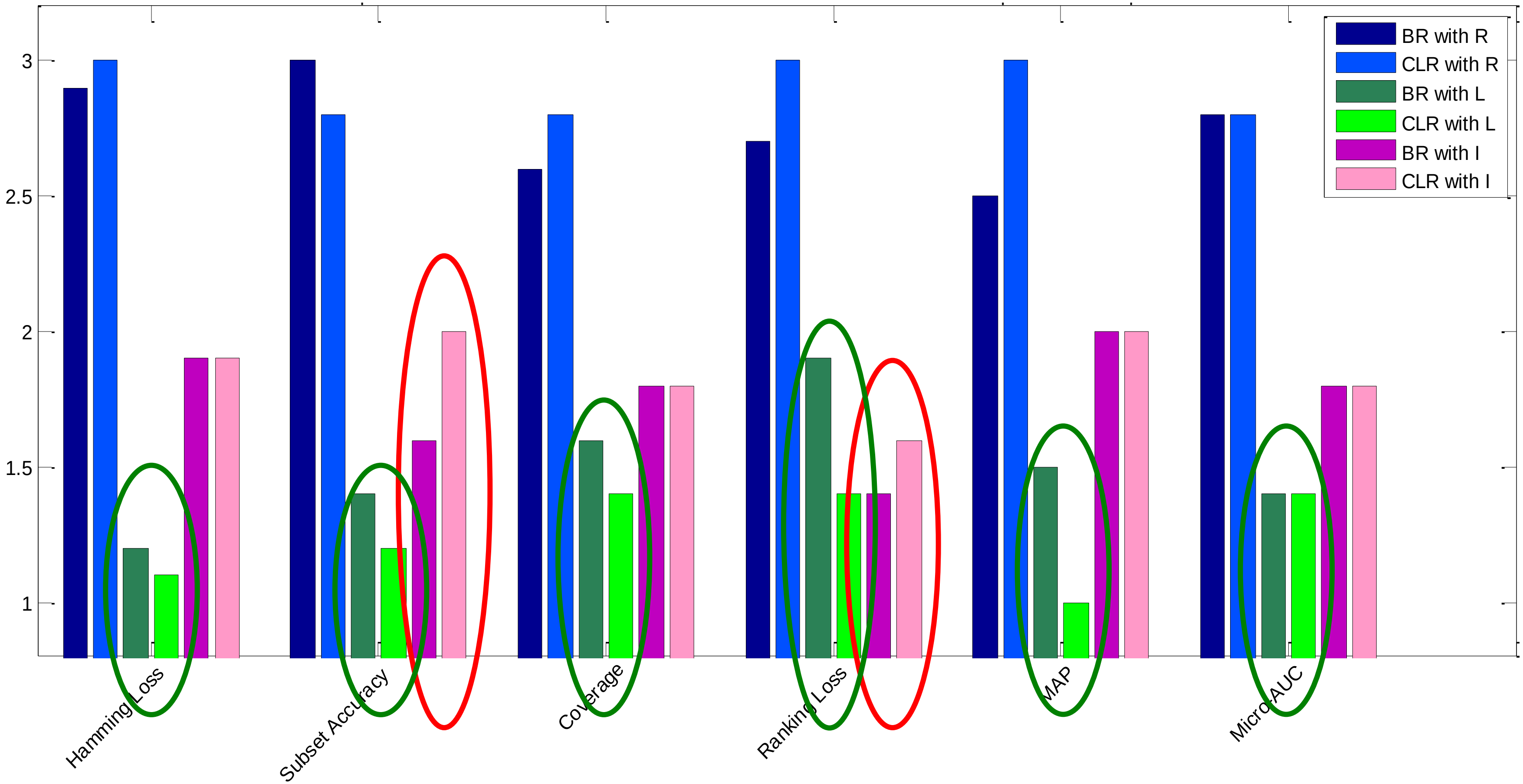
Average Ranking for CLR

- On 5 datasets with #labels < 50 for complexity reasons (those that #labelsets/#examples ≤ 0.1)



BR vs CLR

- On 5 datasets where #labelsets/#examples ≤ 0.1



Iterative stratification suits BR

Labelsets-based suits CLR

Conclusions

- Labelsets-based stratification
  - Works well when #labelsets/#examples is small
  - Works well with Calibrated Label Ranking
- Iterative stratification
  - Works well when #labelsets/#examples is large
  - Works well with Binary Relevance
  - Works well for estimating the Ranking Loss
  - Handles rare labels in a better way
  - Maintains the imbalance ratio of each label in each subset
- Random sampling
  - Is consistently worse and should be avoided, contrary to the typical multi-label experimental setup of the literature

Future Work

- Iterative stratification
  - Investigate the effect of changing the algorithm to respect the desired number of examples at each subset
- Hybrid approach
  - Stratification based on labelsets of the examples of frequent labelsets
  - Iterative stratification for the rest of the examples
- Sampling and generalization performance
  - Conduct statistically valid experiments to assess the quality of the sampling schemes in terms of estimating the test error (unbiased and low variance)