ENHANCING ACCURACY OF MULTILABEL CLASSIFICATION BY EXTRACTING HIERARCHIES

Alexander Ulanov, Nikolay Lyubomishchenko (HP Labs Russia)
German Sapozhnikov and Georgy Shevlyakov (SPbSPU)
31.08.2011

TIR 2011

ABOUT

– Motivation

– Problem

– Hierarchical Multi-Label classification

– State of the art

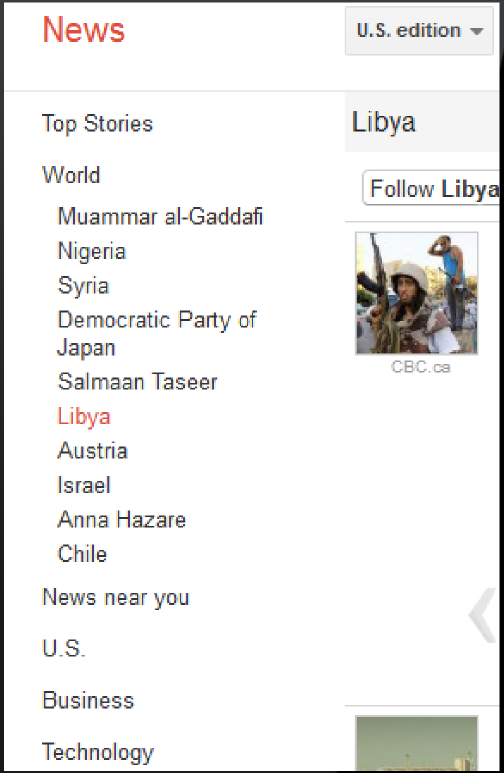
– Our solution, Algorithm

– Experiment results

– Conclusion and future work

MOTIVATION

- Why hierarchical classification can be better:
  - In flat classification, the number of training examples associated with each label is considerably less than the total number of examples
  - The computational complexity of training a multilabel classifier is strongly affected by the number of labels
  - Each classifier in hierarchy deals with much smaller set of labels as compared to L(full set of labels)
- Sometimes datasets already have hierarchy but we are interested in cases when they don't
- Examples of large scale multilabel products:
  - WIPO IPC(International Patent Classification)
  - LexisNexis Hierarchy



WORLD INTELLECTUAL PROPERTY ORGANIZATION

LexisNexis®

TIR 2011

PROBLEM

- **Problem**: Increase performance of large scale multilabel classification
- **Idea:** Transformation of a multilabel classification task with a large set of labels into a tree-shaped hierarchy of simpler multilabel classification tasks

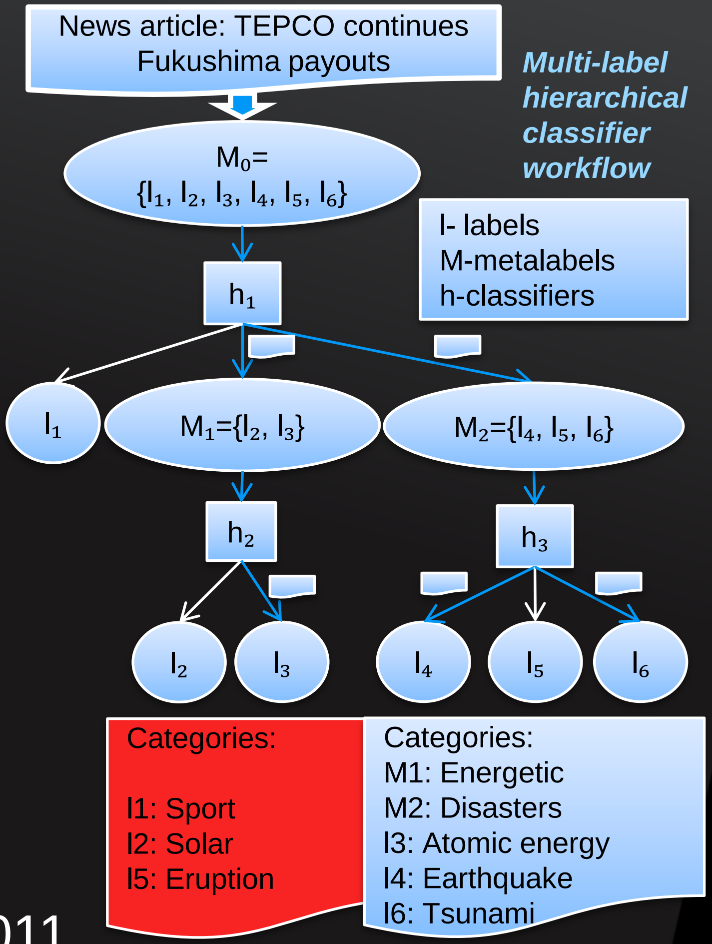
HIERARCHICAL MULTI LABEL CLASSIFICATION

– If number of categories for each document in classification task is more than one, we deal with multi-label classification

– Classes may have the hierarchical structure, we can use it for classification

– On each layer classifier assign document to one or more label

– Each classifier deal with small number of labels



STATE OF THE ART

– Methods that use predefined hierarchy

– Two types of hierarchical classification [Sun et al., 01]
  - Big-bang approach
  - Top-down approach (the most popular)

– HOMER method [Tsoumakas et al., 08] constructs a Hierarchy Of Multi-label classifiERs
  - Automatically organizes labels into a tree-shaped hierarchy
  - Balance clustering algorithm
  - In our algorithm, we use the same concept of hierarchy and metalabels

– [Sapozhnikova et al., 11]
  - Automaticaly extracting hierarchical relationships between classes
  - In our research, we pursue a similar objective.

OUR SOLUTION

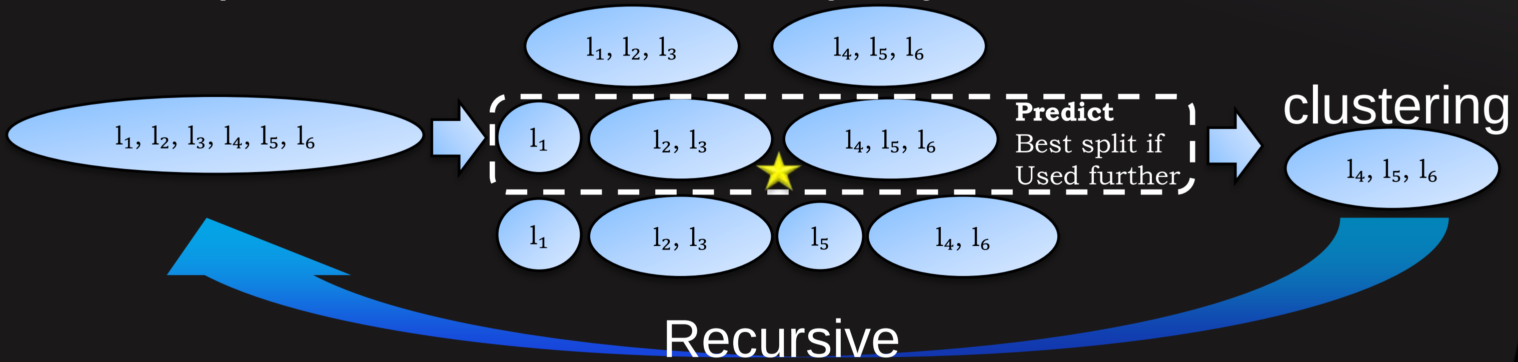
We propose a predictive algorithm for extracting prospective hierarchies:

- Automatic generation of hierarchies for classification using clustering
- Optimize the hierarchy, not classifier technology
- Use criteria that optimizes different measures: precision, recall, or F1
- Toolkit implemented on the basis of Weka ML tool

ALGORITHM: BUILDING TAXONOMIES LAYER BY LAYER

We use top-down approach

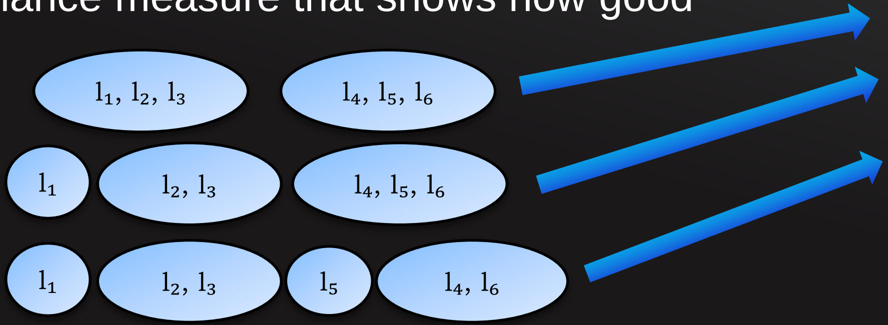
1. We have some number of labels
2. Use clustering on different number of clusters
3. Predict, which partition will be the best for classification, taking into account that we will use clustering for them further (next slide)
4. Make this process recursive for clusters witch size is more than 2. Cluster of size 1 is ready leaf of our hierarchy. Cluster of size 2 will be separated to two clusters anyway



ALGORITHM: PREDICTION FUNCTION (NOW)

Predict, which partition will be the best for classification, taking into account **<u>their result now</u>** and the fact that we will build some number of layers further

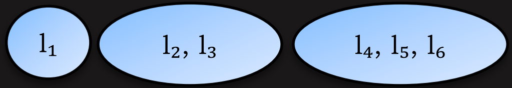
<u>Now</u>: Make classification using clusters as meta-labels. As a result we get a performance measure that shows how good partition are



Number of clusters	Measure
2	0,97
3	0,93
4	0,88

Example: our measure is Accuracy

- We want to predict, how good this partition is



- On this layer Accuracy of the classification on this clusters is 0.93
- To get the prediction for this partition we need to take into account that we will build some number of layers in future (next slide)

TIR 2011

ALGORITHM: PREDICTION FUNCTION (FUTURE)

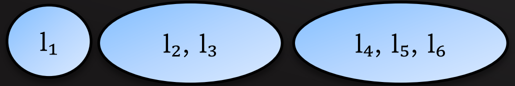
Predict, which partition will be the best for classification, taking into account their result now and **the fact that we will build some number of layers further**

Future: We explore all further possible partitions of cluster and compute the prediction of their performance

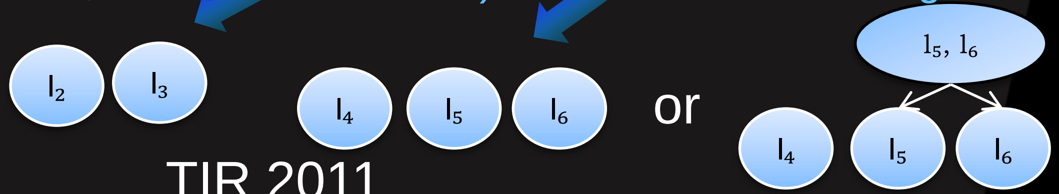
Number of clusters	Measure
2	0,97
3	0,93
4	0,88

Example: our measure is Accuracy

We want to predict, how we can cluster this partition further



1. We need to predict Acc. for $\{l_2, l_3\}$ and $\{l_4, l_5, l_6\}$
2. We believe that for $\{l_2, l_3\}$ it will be the same as in the table (0,97). For $\{l_4, l_5, l_6\}$ prediction can be 0,93 (3 classes into 3 clusters) or 0,97*0,97 (We can divide 3 classes in 2 clusters, size of one will be 2). We will choose the higher value



ALGORITHM: PREDICTION FUNCTION

Predict, which partition will be the best for classification, taking into account their result now and the fact that we will build some number of layers further

Example l_1 l_2, l_3 l_4, l_5, l_6

Number of clusters	Measure
2	0,97
$\boxed{3}$	$\boxed{0,93}$
4	0,88

1. We know the size of all classes and clusters
2. We know the result on this layer
3. We know the prediction for all parts of a partition
   - We made classification and know everything about l_1
   - The predicted Acc. Is 0.97 l_2, l_3
   - The predicted Acc. Is 0.94 (0.97*0.97>0.93) l_4, l_5, l_6
4. For all parts of a partition we calculate prediction of true positives, false negatives….
5. Using all information we can calculate the prediction of Acc. for partition

EXPERIMENTAL RESULTS

Optimizing micro-F1

Dataset name	Classifier typology	Micro			Macro		
		F1	P	R	F1	P	R
Mediamill	flat	0,54	**0,66**	0,45	0,10	**0,24**	0,08
	Hierarchy	0,53	0,58	**0,50**	**0,13**	0,19	**0,11**
Bibtex	Flat	0,31	**0,81**	0,19	0,14	0,40	0,11
	Hierarchy	**0,37**	0,61	**0,27**	**0,22**	0,38	**0,18**
Medical	flat	0,80	0,85	0,75	0,26	0,32	**0,25**
	Hierarchy	0,82	0,84	**0,81**	**0,30**	0,33	0,30
Enron	flat	0,46	**0,66**	0,35	0,09	0,13	0,08
	Hierarchy	**0,50**	0,62	**0,42**	0,10	0,15	0,09

All datasets were taken from http://mulan.sourceforge.net/

– F1 is harmonic mean of precision and recall. It was optimized by optimizing recall at the expense of precision

– We use F1 for each cluster individually (in fact macro-F1) to optimize micro-F1

– We optimized micro-F1 for bibtex and enron. For medical and mediamill result remains the same

– Classifier that we used in each node of the hierarchy: decision tree [Webb.99]

CONCLUSION

– Results of this work:
  - A novel algorithm for automatically extracting hierarchies for classification
  - It is classifier independent. It provides enhancing the accuracy of multilabel classification optimizing a hierarchy structure, not classifiers
  - It intended for datasets which don’t have predefined hierarchies
  - Experimental study of it’s performance on 4 datasets justifies its effectiveness

– Future work:
  - Improve measures for choosing the best layer in the taxonomy
  - If on the next layer prediction comes false we can return on previous layer
  - Make the classifier , that will be good for hierarchy, that we built with our algorithm

Q&A