The Influence of Thresholding Strategy on Multi-label Topic Identification

Jan Lehečka[1]

1 Introduction

The task of multi-label topic identification is to associate each text document in a corpus with one or more relevant label describing the topic (or class, category etc.) of the document. The task is usually solved by a supervised document classifier trained from a set of manually labeled documents. Given an input document, the trained classifier outputs a *soft prediction*, i.e. a vector of scores, one score for each topic, denoting how likely it is the topic is present in the document.

To specify the set of labels to be associated with a document, soft prediction has to be converted to a binary vector (*hard prediction*), with *true* only for few most relevant topics and *false* for others. The set of rules of how to convert a soft prediction into a hard prediction, is called a *thresholding strategy*.

2 Thresholding strategy

A thresholding strategy describes the way how to select a set of relevant labels from a set of K possible labels $\mathscr{L} = \{l_k\}_{k=1}^{K}$ for an arbitrary document d given it's soft prediction . Our approaches are mainly based on learning thresholds from the soft predictions of the training data set, which have been obtained by classifying the training data set after training the classifier, because for this data, we also know the correct hard prediction (true labels).

We tried several approaches to *label-wise thresholding*, where for each label $l_k \in \mathscr{L}$, one threshold t_k is set based on division of obtained scores into two sets:

- S_k^{true} - set of l_k-scores from all documents, where l_k is the true label,
- S_k^{others} - set of l_k-scores from all documents with labels other then l_k.

Then, for each tested document d and for each label l_k, the thresholding strategy is to assign l_k to d if the the score of l_k is higher then:

$$t_k^{true} = \min(S_k^{true}), \tag{1}$$

$$t_k^{others} = \max(S_k^{others}), \tag{2}$$

$$t_k^{mean1} = 0.5[\min(S_k^{true}) + \max(S_k^{others})], \tag{3}$$

$$t_k^{mean2} = 0.5[\text{mean}(S_k^{true}) + \text{mean}(S_k^{others})]. \tag{4}$$

Another strategy we have tried is *sample-wise thresholding*, where the threshold t_d for each document d is obtained from a linear regressor $\mathscr{R}$ trained from the soft predictions of the

[1] student of the doctoral study programme Applied Sciences and Informatics, specialization Cybernetics, e-mail: jlehecka@kky.zcu.cz

training data set. Target values for training the regressor were set in the middle of mean score of document's true labels and mean score belonging to an irrelevant labels (i.e. in the spirit of (4) but in a sample-wise manner). After the regressor $\mathscr{R}$ is trained, it produces a suitable threshold for each document given it's soft prediction. We also tried sorting the scores for each document before training the regressor, i.e. we didn't care which label is relevant for the document d , but we rather trained the regressor from differences between successive scores. We denote these thresholds as $t_d^{\mathscr{R}sort}$.

3 Results

In this paper, we are demonstrating the influence of described thresholding strategies on a multi-label topic identification of Czech news articles in our large web-mined corpus described in Švec at al. (2011). Our training data set consists of $205k$ documents ($70M$ words total) with vocabulary size $700k$ and $21k$ different labels. Because of the lack of training data assigned to low-frequency labels, we used only labels assigned to at least 30 documents, which decreased the number of labels to 1843. Our testing data consists of $44k$ documents.

As a baseline strategy, we used simple and widely used *topN* strategy, also known as *RCut* (rank-based thresholding), which selects N most probable labels. We set N to the average number of labels in the training data set, which was 3.

As a document classifier, we used SVC with linear kernel function and for a data representation, we used sublinear *tf-idf* vector space model. As a metric to measure multli-label topic identification performance, we used sample-wise average precision P, recall R and it's harmonical mean F_1 score.

Table 1: Multi-label topic identification performance with different thresholding strategies

strategy	P	R	F_1
top3 (baseline)	0.759	0.607	**0.655**
t_k^{true}	0.269	0.755	0.376
t_k^{others}	0.784	0.468	0.554
t_k^{mean1}	0.765	0.639	0.668
t_k^{mean2}	0.758	0.672	**0.685**
$t_d^{\mathscr{R}}$	0.811	0.635	0.684
$t_d^{\mathscr{R}sort}$	0.772	0.683	**0.695**

Obtained results are sumarized in Tab. 1, where can be seen, that we can improve (in terms of F_1 score) the performance of multli-label topic identification on our data by 4.58% relatively when using label-wise thresholding (t_k^{mean2}) and roughly the same when using sample-wise thresholding ($t_d^{\mathscr{R}}$). An interesting result is that in the case of sample-wise thresholding, the performance can be further improved just by sorting scores on the input of the trained regressor.

References

Švec, J., Hoidekr, J., Soutner, D., and Vavruška, J., 2011. Web text data mining for building large scale language modelling corpus. *Text, Speech and Dialogue*, pp. 356-363.