FAKULTA
APLIKOVANÝCH VĚD
ZÁPADOČESKÉ
UNIVERZITY
V PLZNI

STUDENTSKÁ
VĚDECKÁ KONFERENCE 2014

The Use of the Unconstrained Cohort Normalization Technique for Multi-label Classification Score Normalization

Lucie Skorkovská[1]

1 Introduction

The goal of the text classification is to categorize a set of documents into predefined set of topic classes or categories. Usually in the field of text classification we are considering only the multiclass classification, where unlike in the binary classification there is more than two possible classes. The simplest task of the text classification is to assign one topic to each document, but in the task of newspaper article topics identification it is especially essential to use the multi-label classification. Its goal is to find a set of labels belonging to each data item. We are using the generative classifier, where the classifier outputs a distribution of probabilities (or likelihood scores), to tackle this task, but the problem with this approach is that the threshold for the positive classification must be set. This threshold can vary for each document depending on the content of the document (words used, length of the document, ...).

The described method for finding a threshold defining the boundary between the "correct" and the "incorrect" topics of a newspaper article is based on the Unconstrained Cohort Normalization (UCN) technique used in the speaker identification task.

2 Score Normalization Technique

For the topic identification we use the multinomial Naive Bayes classifier (Skorkovská et al. (2011)), which outputs a likelihood topic distribution of $p(A|T)$. Now we have to choose the threshold for the selection of the topics to assign to an article. The right way to select the "correct" topics for an article would be setting a dynamic threshold, which should be somehow dependent on the article topic likelihood distribution. A score normalization methods have been used to tackle the problem of the compensation for the distortions in the utterances in the second phase of the open-set text-independent speaker identification problem (Sivakumaran et al. (2003)).

A frequently used form to represent the normalization process is the following:

$$L(A) = \log p(A|T_C) - \log p(A|T_I). \tag{1}$$

where $P(T_C|A)$ is the score given by the correct topic model and $P(T_I|A)$ is the score given by the incorrect topic model. Since the normalization score $\log p(A|T_I)$ of an incorrect topic is not known, it can be approximated by the Unconstrained Cohort model (Auckenthaler et al. (2000)). For every topic model a set (cohort) of N similar models $C = \{T_1, ..., T_N\}$ is chosen. These models in the set C are the most competitive models with the reference topic model, i.e.

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

Table 1: Comparison of different threshold finding methods

metric / method(H)	**3 topics**	**GTMN**	**UCN**
$P(H,D)$	0.5859	0.5916	**0.6650**
$R(H,D)$	0.6155	0.6992	**0.6311**
$F_1(H,D)$	0.6003	0.6409	**0.6476**

models which yield the next N highest likelihood scores. The normalization score is given by:

$$\log p(A|T_I) = \log p(A|T_{UCN}) = \frac{1}{N}\sum_{n=1}^{N} \log p(A|T_n). \qquad (2)$$

Even when we have the topic likelihood score normalized, we still have to set the threshold for verifying the correctness of each topic in the list. Selecting a threshold in a list of normalized likelihoods is more robust, because the normalization removes the influence of the various document characteristics. In our former experiments with score normalization we have defined the threshold as 80$\%$ of the normalized score of the best scoring topic. The topics which achieved better normalized score are the “correct” topics to be assigned. The threshold selected in this way has experimentally proven to be robust, the change in the range of percents does not influence the result of the topic identification. For the UCN normalization, we have chosen the same threshold - 80$\%$ of the best scoring topic, and we have performed experiments with N - size of the set C to be chosen. In the Table 1 the results of the experiments on the collection containing 31k articles is shown.

3 Conclusion

The proposed Unconstrained Cohort Normalization technique achieved 1$\%$ relative improvement compared to the GTMN method and 7.9$\%$ relative improvement compared to the selection of fixed number of topics. Score normalization techniques are very useful in topic identification task, although we still have to set the threshold for verifying the correctness of the topics, selecting a threshold defining the boundary between the correct and the incorrect topics is more robust, because the normalization removes the influence of the various document characteristics.

Acknowledgement

The work has been supported by the grant of The University of West Bohemia, project No. SGS-2013-032.

References

Auckenthaler, R., Carey, M., Lloyd-Thomas, H.: Score normalization for text-independent speaker verification systems. *Digital Signal Processing* 10(13), 42 – 54 (2000)

Sivakumaran, P., Fortuna, J., Ariyaeeinia, M., A.: Score normalisation applied to open-set, text-independent speaker identification. *Proceedings of Eurospeech 2003.* pp. 2669–2672. Geneva (2003)

Skorkovská, L., Ircing, P., Pražák, A., Lehečka, J.: Automatic topic identification for large scale language modeling data filtering. *Text, Speech and Dialogue, LNCS*, vol. 6836, pp. 64–71. Springer Berlin / Heidelberg (2011)