FAKULTA APLIKOVANÝCH VĚD ZÁPADOČESKÉ UNIVERZITY V PLZNI
STUDENTSKÁ VĚDECKÁ KONFERENCE 2013

Multi-label Classification of Newspaper Articles

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.

Two main approaches to the text classification can be identified - the discriminative techniques like support vector machines (Joachims (1998)), decision trees (Schapire et al. (2000)) and neural networks; and generative techniques like Naive Bayes classifier (McCallum (1999)) and Expectation Maximization based methods.

Our experiments regard the field of generative classification, where the classifier outputs a distribution of probabilities (or likelihood scores) and a method for processing this distribution into the sets of the "correct" and the "incorrect" topics is needed. The described method for finding a threshold defining the boundary between the "correct" and the "incorrect" topics of a newspaper article is based on general topic model normalisation.

1.1 General Topic Model Normalisation Method

For the topic identification we use the multinomial Naive Bayes classifier (NBC), chosen due to the results of experiments published in Skorkovská et al. (2011). We have to choose the threshold for the selection of the topics to assign to an article. So far we have been selecting the best 3 topics for each article. This is not the best way, because some short articles can concern only one topic, on the other hand some long articles, especially from the politics category often incorporate many other topics. 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.

The *General topic model normalisation (GTMN)* method for finding the threshold is inspired by the Universal background model (UBM) normalisation technique used in the speaker recognition task (Sivakumaran et al. (2003)). First, the NBC classifier is used to output a likelihood topic distribution. Then, the topic likelihood scores $\hat{P}(T|A)$ are normalised with the score of the general model (created as a language model of the whole collection) $\hat{P}(G|A)$:

$$\hat{P}(T|A)_{GTMN} = \frac{\hat{P}(T|A)}{\hat{P}(G|A)} \tag{1}$$

Now we have a list of the likelihoods normalised by the general topic model, specifically

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

we have the list of how better the topics describe the article in comparison with the general topic model. We select only the topics which are better scoring than the general topic model and we make the assumption that the topics which have at least 80 percent of the normalised score of the best scoring topic are the "correct" topics to be assigned.

For the experiments the collection containing 31 419 articles was used (Skorkovská (2012)). In the Table 1 the General topic model normalisation method for finding the threshold is compared to the previously used selection of 3 topics for each article.

Table 1: Comparison of different threshold finding methods

metric / method(H)	**3 topics**	**GTMN**
$Precision(H, D)$	0.5859	**0.5916**
$Recall(H, D)$	0.6155	**0.6992**
$F_1 - measure(H, D)$	0.6003	**0.6409**

2 Conclusion

The GTMN method achieved better results than the previously used selection of 3 topics. The 80 percent threshold was found out experimentally, but we discovered that after the GTMN there is a huge difference between the "correct" and the "incorrect" topic scores, therefore setting the threshold is not sensitive. In the future work, we will test the method on other collections with different number of topic categories to confirm the universality of this method.

Acknowledgement

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

References

Schapire, R.E., Singer, Y.: Boostexter: A boosting-based system for text categorization. *Machine Learning.* pp. 135–168 (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.: Application of lemmatization and summarization methods in topic identification module for large scale language modeling data filtering. *Text, Speech and Dialogue, LNCS*, vol. 7499, pp. 191–198. Springer Berlin Heidelberg (2012)

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)

McCallum, A.K.: Multi-label text classification with a mixture model trained by em. *AAAI 99 Workshop on Text Learning* (1999)

Joachims, T.: Text categorization with support vector machines: Learning with many relevant features. *Machine Learning: ECML-98, LNCS*, vol. 1398, pp. 137–142. Springer, Heidelberg (1998)