Score Normalization Methods Applied to Topic Identification*

Lucie Skorkovská, Zbyněk Zajíc

University of West Bohemia, Faculty of Applied Sciences
New Technologies for the Information Society
Univerzitní 22, 306 14 Plzeň, Czech Republic
{lskorkov,zzajic}@ntis.zcu.cz

Abstract. Multi-label classification plays the key role in modern categorization systems. Its goal is to find a set of labels belonging to each data item. In the multi-label document classification unlike in the multi-class classification, where only the best topic is chosen, the classifier must decide if a document does or does not belong to each topic from the predefined topic set. We are using the generative classifier 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, ...). In this paper we use the Unconstrained Cohort Normalization, primary proposed for speaker identification/verification task, for robustly finding the threshold defining the boundary between the "correct" and the "incorrect" topics of a document. In our former experiments we have proposed a method for finding this threshold inspired by another normalization technique called World Model score normalization. Comparison of these normalization methods has shown that better results can be achieved from the Unconstrained Cohort Normalization.

Keywords: topic identification, multi-label text classification, Naive Bayes classification, score normalization

1 Introduction

Multi-label classification is increasingly required in modern categorization systems, especially in the fields of newspaper article topic identification, social network comments classification, web content topical organization or email routing, where each "document" (either newspaper article or email) can belong to many topics (or keywords or tags) selected from a large set of possible labels. Usually, the multi-label classification is handled through a set of binary classifiers, one for each label, deciding if a document does or does not belong to a specified topic. The issue with this approach is that for each topic the classifier must be trained and the threshold for the positive classification must be set. This may not be a problem for a classification task with a small set of topics (ten topics for example), where for each one of them a sufficient amount of training data is

* The work was supported by the Ministry of Education, Youth and Sports of the Czech Republic project No. LM2010013 and University of West Bohemia, project No. SGS-2013-032.

available, but in a real application the set of topics is usually quite large (450 topics in our case) and for some of them very little training data can be obtained.

Possible alternative is to use a single generative classifier like Naive Bayes (NB) classifier [1][8], which outputs a distribution of probabilities (or likelihood scores) of the document belonging to the topics from the topic set. In this approach only a single threshold defining the boundary between the "correct" and the "incorrect" topics of a document has to be set. The problem addressed in this paper is how to process the distribution of topics and select this threshold. Since it may vary depending on the content of each document, it can not be fixed for the whole document collection, but a dynamically set threshold is needed.

In our former experiments we have proposed a General Topic Model Normalization (GTMN) method [14] for finding the threshold inspired by the World Model score normalization technique used in speaker identification/verification task. Since this method has shown promising results, in this paper we try to propose advanced technique for the threshold selection based on another technique used in speaker identification area - Unconstrained Cohort Normalization (UCN).

The score normalization methods are used to improve the newspaper topic identification results in a real-life application for language modeling data filtering [17], where the topics are chosen from a quite extensive hierarchy - it contains about 450 topics.

2 Multi-label Text Classification

The multi-label classification methods can be divided into two main categories - *data transformation methods* and *algorithm adaptation methods*. The methods of the first group transform the problem into the single-label classification problem and the methods in the second group extend the existing algorithms to handle the multi-label data directly. In [16] a detailed overview of the existing *data transformation methods* is presented: The easiest way is to transform the multi-label data set into single-label by either selecting only one label from the multiple labels for each data instance or by discarding every multi-label data instance from the set. Another option is to considers each set of labels as one label together [8].

The most common option is to train a binary one-vs.-rest classifier for each class. The labels for which the binary classifier yields a positive result are then assigned to the tested data item. The disadvantage of this method is that you have to transform the data set into $|L|$ data sets, where L is the set of possible labels, containing only the positive and negative examples. The second disadvantage is that you have to find the threshold for each binary classifier. This method was used for example in [4][18].

Another possibility is to decompose each training data with n labels into n data items each with only one label. One generative classifier with the distribution of likelihoods for all labels is learned from the transformed data set. The distribution is then processed to find the correct labels of the data item. This approach is used in the work [3][8] and also in our experiments.

2.1 Threshold Definition for Generative Classifiers

A related work on the problem how to select the set of correct topics from the output distribution of the generative classifiers is presented in this section. A straightforward approach is to select the labels for which the likelihood is greater than a specific threshold or select a predefined number of topics. In the work [1] only the one best label is assigned to each news article. In our later work, we selected 3 topics for each article [15]. In the work [8] this problem is bypassed by creating a mixture topic model from all possible topic subsets and then choosing the subset for which the corresponding mixture model has achieved the maximum likelihood.

To our knowledge, the only work concerning the finding of a threshold for choosing the correct topics in the distribution output of a classifier is described in [3]. A dynamic threshold is set as the mean plus one standard deviation (MpSD) of the topic likelihoods. The assumption is that topics that have a likelihood greater than this threshold are the best choices for the article. In our former experiments [14] we have proposed a General Topic Model Normalization (GTMN) method for finding the threshold inspired by the World Model score normalization technique and compared it to the related methods. The results obtained from the comparison can be seen in Section 4.4 in Table 1.

3 Score Normalization Applied to Multi-label Topic Identification

The topic identification problem is quite similar to the open-set text-independent speaker identification (OSTI-SI) problem. Similarly as in the speaker identification, the multi-label document classification can be described as a twofold problem: First, the speaker model best matching the utterance has to be found and secondly it has to be decided, if the utterance has really been produced by this best-matching model or by some other speaker outside the set. The difficulty in this task is that the speakers are not obliged to provide the same utterance that was the system trained on. The document classification problem can be described in the same way: First, we need to find the topic models which have the best likelihood score for the tested document and second, we have to choose only the correct topic models which really generated the document. The only difference in topic identification is that we try to find more than one correct topic model. The normalization methods from OSTI-SI can be used in the same way, but have to be applied to all topic models likelihoods.

3.1 Naive Bayes Classification

For the first phase of the topic identification the multinomial Naive Bayes classifier is used, which is formally equal to the language modeling based approach in the information retrieval [7]. Each topic is defined by its unigram language model and a probability of a document A being generated by a topic model T is a conditional model $P(T|A)$. Using the Bayes' theorem and leaving out the prior probability of an article $P(A)$, the following equation can be written:

$$P(T|A) \propto \frac{P(T)p(A|T)}{P(A)} \propto p(A|T), \tag{1}$$

where $P(T)$ is the prior probability of a topic T, which can be estimated as a relative frequency of a topic in training data, or considered uniform and be left out as in our case [14]. The distribution of topic likelihoods $p(A|T)$ is then used to find the most likely topics of an article. Under the "naive" conditional independence assumption $p(A|T)$ can be computed in the following way:

$$p(A|T) = \prod_{t \in A} p(t|T), \qquad \hat{p}(t|T) = \frac{tf_{t,T}}{N_T}, \tag{2}$$

where $p(t|T)$ is a conditional probability of a term t given the topic T. This probability is estimated by the maximum likelihood estimate as the relative frequency of the term t in the training data of the topic T. The uniform prior smoothing was used in the estimation of $p(t|T)$.

3.2 Score Normalization

Now that we have the distribution of topic likelihoods $p(A|T)$ we have to find the threshold for selection the correct topics of an article. 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 [12]. In the topic identification task, the likelihood score of a topic obtained from the classifier is dependent on the characteristics of the document (words used, length of the document, ...).

Similarly as in the OSTI-SI [12] we can define the decision formula:

$$P(T_C|A) > P(T_I|A) \rightarrow A \in T_C \quad \text{else} \quad A \in T_I, \tag{3}$$

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. By the application of the Bayes' theorem, formula (3) can be rewritten as:

$$\frac{p(A|T_C)}{p(A|T_I)} > \frac{P(T_I)}{P(T_C)} \rightarrow A \in T_C \quad \text{else} \quad A \in T_I, \tag{4}$$

where $l(A) = \frac{p(A|T_C)}{p(A|T_I)}$ is the normalized likelihood score and $\theta = \frac{P(T_I)}{P(T_C)}$ is a threshold that has to be determined. Setting a threshold θ a priori is a difficult task, since we do not know the prior probabilities $P(T_I)$ and $P(T_C)$. Similarly as in the OSTI-SI the topic set is open - an article belonging to a topic not contained in our set can easily occur.

A frequently used form to represent the normalization process is the following [12]:

$$L(A) = \log p(A|T_C) - \log p(A|T_I). \tag{5}$$

The score $\log p(A|T_C)$ is affected by the document characteristics as well as the score $\log p(A|T_I)$. Thus, the distance between them should stand constant for various documents and finding the threshold experimentally for the whole collection of documents can be achieved.

Since the normalization score $\log p(A|T_I)$ of an incorrect topic is not known, there are several possibilities how to approximate it:

General Topic Model Normalization (GTMN) The unknown model T_I can be approximated by the General topic model G [14] which was created as a language model from all documents in the training collection. This technique was inspired by the World Model normalization [11]. The normalization score of a topic model T_I is defined as:

$$\log p(A|T_I) = \log p(A|G) \tag{6}$$

Unconstrained Cohort Normalization (UCN) In this method [2], 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. models which yield the next N highest likelihood scores. The normalization score is given by:

$$\log p(A|T_I) = \log p(A|T_{UCN}) = \frac{1}{N} \sum_{n=1}^{N} \log p(A|T_n). \tag{7}$$

Even when we have the topic likelihood score normalized, we still have to set the threshold θ in (4) for verifying the correctness of each topic in the list. Selecting a threshold defining the boundary between the correct and the incorrect topics in a list of normalized likelihood is more robust, because the normalization removes the influence of the various document characteristics. In our former experiments with GTMN [14] we have selected only the topics which are better scoring than the general topic model and 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.

4 Performed Experiments

In this section the experiments with the UCN score normalization method are presented. All experiments were performed with the topic identification module which is a part of the System for acquisition and storing data [17] designed to gather the training data for the estimation of the parameters of statistical language models for natural language processing. For the topic identification experiments the most important parts of the system are the text preprocessing modules. On each article a *tokenization*, *text normalization*, *vocabulary-based substitution* and *decapitalization* algorithms are applied. Automatic *text lemmatization* [6] is also applied in our work, since it has been shown to improve the results when dealing with sparse data [5] [10] in highly inflected languages.

4.1 Topic Identification Module

Since it has been shown that not only the size of the training data is important, but also the right scope of the language models training texts is needed [9], the topic identification algorithm is used for large scale language modeling data filtering [15]. The

topic identification module uses a multinomial Naive Bayes classifier, since based on the nature of our application (every day more than 600 new articles are downloaded containing more than 130 new topic training articles) we needed the topic identification algorithm which will be fast and can use the easily updatable statistics stored in the database tables as the trained classifier data. The motivation for choosing the NB classifier is more addressed in [14][15].

The topics are chosen from a hierarchical system, which is built in a form of a topic tree and is based on our expert findings in topic distribution in the articles on the Czech favorite news servers. Totally it contains about 450 topics and topic categories. The advantage of the hierarchical organization of the topics is currently used only for the selection of documents to be used as the training data for the estimation of the parameters of statistical language models but not for the topic identification. For the classification all topic are used only as the set of topics on an equal level (all 450 topics). This is caused by the nature of the training data since we use as training data the real articles from the different news servers and we do not want to change it in any way. The authors of these articles to our knowledge do not use any topic hierarchy, or at least not strictly and easily readable from the data. Sometimes the articles have assigned also the more general topic for some detailed topic, but mostly it does not (for example the article with the topic `soccer` mostly does not have also the topic `sports`.

4.2 Evaluation Metrics

The evaluation of the results of the multi-label classification requires different metrics than those used in the single-label classification. The commonly used metric is somewhat similar to the evaluation used in the field of information retrieval (IR), where each article undergoing classification is considered to be a query in IR and precision and recall is computed for the answer topic set. Similar measures was used in [4]. For the article set D and the classifier H precision ($P(H, D)$) and recall ($R(H, D)$) is computed:

$$P(H,D) = \frac{1}{|D|}\sum_{i=1}^{|D|}\frac{T_C}{T_A}, \qquad R(H,D) = \frac{1}{|D|}\sum_{i=1}^{|D|}\frac{T_C}{T_R}, \tag{8}$$

where T_A is the number of topics assigned to the article, T_C is the number of correctly assigned topics and T_R is the number of relevant reference topics. The $F_1(H, D)$-measure, which is used for straightforward comparison of methods, is then computed from the $P(H, D)$ and $R(H, D)$ measures:

$$F_1(H,D) = 2\frac{P(H,D)\cdot R(H,D)}{P(H,D)+R(H,D)}. \tag{9}$$

These metrics used for the evaluation of multi-label classification express also the partial match of the classification result, so they have to be understood slightly different than those used in single-label classification. For each data item being classified (news article in this case) we obtain a precision and a recall values expressed as a percentage of the full match between the correct topics set and the assigned topics set. The metrics computed for the whole set of articles ($P(H, D)$, $R(H, D)$ and $F_1(H, D)$-measure)

therefore express how the classification of an article is "good" on average (e.g. the result $F_1(H, D) = 0.66$ means that on average the classification of an article is 66% good - 2 correct and 1 incorrect topics was assigned to an article with 3 relevant topics).

4.3 Test Data

For the experiments a smaller collection containing 31 419 articles from the news server *ČeskéNoviny.cz* separated from the whole corpus was used [13]. The collection contains articles published in the year 2011(January to October) and is divided into 27 000 training and 4 419 testing articles. The articles have not been rearranged in any way, therefore all the test articles was published after the training articles and may contain events not described in the training set. This reflects the real situation in our system, where we need to identify the topics of each newly downloaded article.

Table 1. Comparison of different threshold finding methods

metric / method(H)	1 topic	3 topics	MpSD	GTMN	UCN
$P(H, D)$	0.8123	0.5859	0.0554	0.5916	**0.6650**
$R(H, D)$	0.3191	0.6155	0.9611	0.6992	**0.6311**
$F_1(H, D)$	0.4582	0.6003	0.1048	0.6409	**0.6476**

4.4 Results

The results of the UCN method applied to the topic identification score for robustly finding the threshold are compared to the results of the GTMN method proposed in [14]. The results are also compared to the previously used selection of 3 topics for each article and also to selection only one topic [1] and setting the threshold as the mean plus one standard deviation (MpSD) of the topic likelihoods [3]. For UCN the size of the cohort was selected experimentally $N = 80$. The experiments (see Table 1) with score normalization techniques from speaker identification domain has shown significantly better results than other techniques used for threshold selection in multi-label document classification. The newly proposed UCN method yields even better results than previously tested GTMN method.

5 Conclusions

This article has proved that score normalization techniques are very useful in topic identification task. The score normalization methods are not time consuming, therefore they can be used even in real-life application like ours. 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. The proposed UCN technique achieved 1% relative improvement compared to the GTMN

method and 7.9% relative improvement compared to the selection of fixed number of topics. The same evaluation was repeated on a different collection of documents separated from our database and the results has shown the same trend.

References

1. Asy'arie, A.D., Pribadi, A.W.: Automatic news articles classification in indonesian language by using naive bayes classifier method. In: Proc. of the 11th Int. Conf. iiWAS2009. pp. 658–662. ACM, New York, USA (2009)
2. Auckenthaler, R., Carey, M., Lloyd-Thomas, H.: Score normalization for text-independent speaker verification systems. Digital Signal Processing 10(13), 42 – 54 (2000)
3. Bracewell, D.B., Yan, J., Ren, F., Kuroiwa, S.: Category classification and topic discovery of japanese and english news articles. Electron. Notes Theor. Comput. Sci. 225, 51–65 (2009)
4. Godbole, S., Sarawagi, S.: Discriminative methods for multi-labeled classification. In: Proc. of PAKDD2004. pp. 22–30. Springer (2004)
5. Ircing, P., Müller, L.: Benefit of Proper Language Processing for Czech Speech Retrieval in the CL-SR Task at CLEF 2006. In: Evaluation of Multilingual and Multi-modal Information Retrieval - 7th Workshop of the CLEF. pp. 759–765. LNCS, Alicante, Spain (2007)
6. Kanis, J., Müller, L.: Automatic lemmatizer construction with focus on oov words lemmatization. In: TSD 2005, LNCS, vol. 3658, pp. 742–742. Springer Berlin (2005)
7. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to Information Retrieval. Cambridge University Press, New York (2008)
8. McCallum, A.K.: Multi-label text classification with a mixture model trained by em. In: AAAI 99 Workshop on Text Learning (1999)
9. Psutka, J., Ircing, P., Psutka, J.V., Radová, V., Byrne, W., Hajič, J., Mírovský, J., Gustman, S.: Large vocabulary ASR for spontaneous Czech in the MALACH project. In: Proceedings of Eurospeech 2003. pp. 1821–1824. Geneva (2003)
10. Psutka, J., Švec, J., Psutka, J.V., Vaněk, J., Pražák, A., Šmídl, L., Ircing, P.: System for fast lexical and phonetic spoken term detection in a czech cultural heritage archive. EURASIP J. Audio, Speech and Music Processing 2011 (2011)
11. Reynolds, D.A., Quatieri, T.F., Dunn, R.B.: Speaker verification using adapted gaussian mixture models. In: Digital Signal Processing. p. 2000 (2000)
12. Sivakumaran, P., Fortuna, J., Ariyaeeinia, M., A.: Score normalisation applied to open-set, text-independent speaker identification. In: Eurospeech. pp. 2669–2672. Geneva (2003)
13. Skorkovská, L.: Application of lemmatization and summarization methods in topic identification module for large scale language modeling data filtering. In: TSD 2012, LNCS, vol. 7499, pp. 191–198. Springer Berlin (2012)
14. Skorkovská, L.: Dynamic threshold selection method for multi-label newspaper topic identification. In: TSD 2013, LNCS, vol. 8082, pp. 209–216. Springer Berlin (2013)
15. Skorkovská, L., Ircing, P., Pražák, A., Lehečka, J.: Automatic topic identification for large scale language modeling data filtering. In: TSD 2011, LNCS, vol. 6836, pp. 64–71. Springer Berlin (2011)
16. Tsoumakas, G., Katakis, I.: Multi-label classification: An overview. Int J Data Warehousing and Mining 2007, 1–13 (2007)
17. Švec, J., Hoidekr, J., Soutner, D., Vavruška, J.: Web text data mining for building large scale language modelling corpus. In: TSD 2011, LNCS, vol. 6836, pp. 356–363. Springer Berlin (2011)
18. Zhang, M.L., Zhou, Z.H.: A k-nearest neighbor based algorithm for multi-label classification. In: Granular Computing, 2005 IEEE International Conference on. vol. 2, pp. 718–721 (2005)