Score Normalization Methods Applied to Topic Identification*

Lucie Skorkovská and 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 correc 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 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.

* 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.

P. Sojka et al. (Eds.): TSD 2014, LNAI 8655, pp. 133–140, 2014.

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