Proximity based one-class classification with Common N-Gram dissimilarity for authorship verification task

Notebook for PAN at CLEF 2013

Magdalena Jankowska, Vlado Kešelj, and Evangelos Milios

Faculty of Computer Science, Dalhousie University
{jankowsk, vlado, eem}@cs.dal.ca

Abstract We describe our participation in the Author Identification task of the PAN 2013 competition. This competition task presents participants with a set of authorship verification problems. In each such a problem, one is given a set of documents written by one author and a sample document; the task is to answer the question whether or not the sample document was written by the same author as the remaining documents. We approach this problem by proposing a proximity based method for one-class classification (based on an idea similar to the k-center boundary method) that applies the Common N-Gram (CNG) dissimilarity measure. The CNG dissimilarity is based on the differences in the frequencies of the character n-grams that are most common in the considered documents. Our method compares the dissimilarity between the sample document and each document from the target set of documents of known authorship to the maximum dissimilarity between this target document and all other documents from the set; thresholding is applied to arrive at the classification of the sample document. Our method yielded F_1 of 0.659 on the whole competition test dataset and the competition ranking 5th (shared) of 18 (according to the results announced on June 12, 2013).

1 Introduction

Authorship verification problem is a type of authorship attribution problem, in which given a set of documents written by one author, and a sample document, one is to answer the question whether or not the sample document was written by the same author as the remaining documents. The PAN 2013 competition Authorship Identification task provides a testbed for the authorship verification solutions. The test dataset consists of authorship verification problems for text documents in English, Greek and Spanish.

We approach this task with an algorithm based on the idea of proximity based methods for one-class classification, (similar to the idea of the k-center boundary algorithm) that applies the Common N-Gram (CNG) dissimilarity measure.

2 Methodology

Our algorithmic approach for the one-class classification is proximity based and it resembles the idea of the k-centre algorithm for one-class classification [9], [8], with k

being equal to the number of all documents in the target set (i.e., written by the given author). The k-center algorithm uses equal radius sphere boundaries around the target documents and compares the sample document to the closest target document; we propose a different classification condition based on the comparison for each target document the maximum dissimilarity between it and any other target document to the dissimilarity between it and the sample document.

Let $A = \{d_1, ..., d_k\}$, $k \geq 2$, be a set of documents written by a given author (we will later describe how we deal with a situation when only one such document is provided). We will call these documents *target documents*. Let u be a sample document which authorship we are to verify, that is to classify it as either belonging to the *target class* (written by the same author as the documents from A) or not.

Our algorithm calculates for each target document d_i the maximum dissimilarity between this document and all other target documents $D^{max}(d_i, A)$ as well as the dissimilarity between this document and the sample document $D(d_i, u)$, and finally the dissimilarity ratio $r(d_i, u, A) = \frac{D(d_i,u)}{D^{max}(d_i,A)}$. (thus $r(d_i, u, A) < 1$ means that there exists in the target set a document more dissimilar to d_i than u, while $r(d_i, u, A) > 1$ means that all the target documents are more similar to d_i than u). As the measure of dissimilarity between the sample document u and the entire target set A we take the average of the dissimilarity ratio: $M(u, A) = \frac{\sum_{i=1,...,k} r(d_i,u,A)}{k}$. We apply a threshold θ on the value of $M(u, A)$ and classify u as belonging to the target class iff $M(u, A) <= \theta$.

Notice that the dissimilarity between the documents does not need to be an l_2 distance, not even a metric distance (i.e., does not need to fulfil the triangle inequality), as is in fact the case for the dissimilarity measure we have chosen.

For the dissimilarity measure between documents we use the Common N-Gram (CNG) dissimilarity, proposed by Kešelj et al. [5]. It is based on the differences in the in usage frequencies of the most common character n-grams of the considered documents. For each document a sequence of the most common character n-grams coupled with their frequencies (normalized by the length of the document) is extracted; such a sequence is called a *profile* of the document. The dissimilarity between two documents of the profiles P_1 and P_2 is defined as follows:

$$D(P_1, P_2) = \sum_{x \in (P_1 \cup P_2)} \left(\frac{f_{P_1}(x) - f_{P_2}(x)}{\frac{f_{P_1}(x) + f_{P_2}(x)}{2}} \right)^2 \quad (1)$$

where x is a character n-gram from the union of two profiles, and $f_{P_i}(x)$ is the normalized frequency of the n-gram x in the the profile P_i, $i = 1, 2$ ($f_{P_i}(x) = 0$ whenever x does not appear in the profile P_i).

The important parameters of the dissimilarity is the length of the character n-grams n and the length of the profile L.

The CNG dissimilarity (or its variants) applied in a k-Nearest Neighbour classification scheme (Common N-gram classifier) were successfully applied to the authorship classification tasks [5], [3], [6].

In our software we used n-grams in which tokens are utf8-encoded characters. The package Text::Ngrams [4] was used to extract the n-grams and their frequencies. To

select the three parameters: n (length of the character n-grams), L (length of the profile) and θ (threshold for the average dissimilarity ratio), we performed experiments on training datasets of authorship verification in English and Greek, with the objective to maximize the accuracy. We used the training dataset provided for the PAN 2013 Authorship Identification task [2] as well as two other datasets which we compiled using existing datasets for other authorship classification tasks, namely the corpus for the Traditional Authorship Attribution subtask of the the PAN 2012 competition [1] (in English) and the modern Greek dataset B created by Stamatos et al. [7]. Table 1 presents the parameters values we selected for the competition (for Spanish we used the same parameters as for English).

	English and Spanish	Greek
n (n-gram length)	6	7
L (profile length)	2000	2000
θ (threshold) if at least two taget documents are given	1.02	1.008
θ (threshold) if a single target document is given	1.06	1.04

Table 1. The parameters of our method used in the competition.

Our method requires at least two target documents. In cases when only one target document is provided, we split it exactly in half to create two documents. As in this case these two documents are most likely very similar to each other as they originate from a single document, we performed additional experiments on our training datasets (for the previously selected values of n and L) for the cases with a single target document, to arrive at somewhat higher values of the threshold θ for such a case, reported in Table 1.

As our method is based on the ratios of dissimilarities between documents, we took care that the documents in a given problem are always represented by profiles of the same length (by adding a condition that if a profile of a given length cannot be created for some documents within a given problem instance because there is not enough distinct character n-grams in the documents, then the length of all profiles in the instance is shortened accordingly). Similarly, we found out that cutting all documents in a given problem instance to the length of the shortest document tend to increase the accuracy of the method, so we applied this prepossessing.

As our method uses the ranking value to which a threshold is applied, it is possible to represent this value as a confidence score in the range from 0 (corresponding to classifying as not belonging to the target class) to 1 (corresponding to classifying as belonging to the target class) to provide them as part of the answers in the competition. To calculate such confidence scores we linearly scaled the average dissimilarity ratio $M(u, A)$ using the threshold θ, so that the value of the average dissimilarity ratio equal to θ corresponds to the score 0.5, values greater than θ correspond to the scores between 0 and 0.5, and values less than θ correspond to the scores between 0.5 and 1 (a cutoff of 0.1 is applied, i.e. all values of $M(u, A) < \theta - cutoff$ are mapped to the score 1, and all values of $M(u, A) > \theta + cutoff$ are mapped to the score 0).

3 Results

In the PAN 2013 competition task Author Identification our method yielded the results presented in Table 2 (according to the results announced on June 12, 2013). As in the competition it was possible to withdraw an answer for a problem, the recall and precision are defined as follows: recall = #correct_answers/#problems, precision = #correct_answers/#answers. The F_1 measure is the harmonic mean of the precision and recall. As we provided the answers for all problems, in our case the F_1 measure is equivalent to the accuracy, i.e., to the fraction of all problems that have been correctly classified.

	All test data	English test data	Greek test data	Spanish test data
F_1	0.659	0.733	0.600	0.640
competition rank	5th (shared) of 18	5th (shared) of 18	7th (shared) of 16	9th of 16

Table 2. The results in the PAN 2013 competition task Author Identification, according to the results announced on June 12, 2013.

References

1. Pan 2012, task: Author identification. http://www.uni-weimar.de/medien/webis/research/events/pan-12/pan12-web/authorship.html (accessed on Apr 2, 2013)
2. Pan 2013, task: Author identification. http://www.uni-weimar.de/medien/webis/research/events/pan-13/pan13-web/author-identification.html (accessed on Feb 19, 2013)
3. Juola, P.: Authorship attribution. Found. Trends Inf. Retr. 1(3), 233–334 (Dec 2006)
4. Kešelj, V.: Perl Package Text::Ngrams. http://www.cs.dal.ca/ vlado/srcperl/Ngrams (accessed on Feb 1, 2012)
5. Kešelj, V., Peng, F., Cercone, N., Thomas, C.: N-gram-based author profiles for authorship attribution. In: Proceedings of the Conference Pacific Association for Computational Linguistics, PACLING'03. pp. 255–264. Dalhousie University, Halifax, Nova Scotia, Canada (August 2003)
6. Stamatatos, E.: Author identification using imbalanced and limited training texts. In: Proceeding of the 18th International Workshop on Database and Expert Systems Applications, DEXA'07. pp. 237–241 (September 2007)
7. Stamatatos, E., Kokkinakis, G., Fakotakis, N.: Automatic text categorization in terms of genre and author. Comput. Linguist. 26(4), 471–495 (Dec 2000)
8. Tax, D.: One Class Classification. Ph.D. thesis, Delft University of Technology (2001)
9. Ypma, A., Ypma, E., Duin, R.P.: Support objects for domain approximation. In: Proceedings of Int. Conf. on Artificial Neural Networks. pp. 2–4 (1998)