The Class Imbalance Problem in Author Identification

Efstathios Stamatatos
University of the Aegean
stamatatos@aegean.gr

Abstract

Author identification can be seen as a single-label multi-class text categorization problem. Very often, there are extremely few training texts at least for some of the candidate authors or there is a significant variation in the text-length among the available training texts of the candidate authors. Moreover, in this task usually there is no similarity between the distribution of training and test texts over the classes, that is, a basic assumption of inductive learning does not apply. Previous work [3] provided solutions to this problem for *instance-based* author identification approaches (i.e., each training text is considered a separate training instance). This work [4] deals with the class imbalance problem in *profile-based* author identification approaches (i.e., a profile is extracted from all the training texts per author). In particular, a variation of the Common N-Grams (CNG) method, a language-independent profile-based approach [2] with good results in many author identification experiments so far [1], is presented based on new distance measures that are quite stable for large profile length values. Special emphasis is given to the degree upon which the effectiveness of the method is affected by the available training text samples per author. Experiments based on text samples on the same topic from the Reuters Corpus Volume 1 are presented using both balanced and imbalanced training corpora. The results show that CNG with the proposed distance measures is more accurate when only limited training text samples are available, at least for some of the candidate authors, a realistic condition in author identification problems.

References

[1] Juola, P. “Ad-hoc Authorship Attribution Competition”. *Proc. of ALLC/ACH Joint Conf.*, pp. 175-176, 2004.

[2] Keselj, V., F. Peng, N. Cercone, and C. Thomas, “N-gram-based Author Profiles for Authorship Attribution”. *Proc. of the Conf. of Pacific Association for Computational Linguistics*, 2003.

[3] Stamatatos, E., “Text Sampling and Re-sampling for Imbalanced Author Identification Cases”, In *Proc. of the 17th European Conference on Artificial Intelligence (ECAI'06)*, 2006.

[4] Stamatatos, E. “Author Identification Using Imbalanced and Limited Training Texts”, In *Proc. of the 4th International Workshop on Text-based Information Retrieval,* 2007.