Text Sampling and Re-sampling for Imbalanced Authorship Identification Cases

Efstathios Stamatatos[1]

Abstract. Authorship 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. In this paper, we present methods to handle imbalanced multi-class textual datasets. The main idea is to segment the training texts into sub-samples according to the size of the class. Hence, minority classes can be segmented into many short samples and majority classes into less and longer samples. Moreover, we explore text re-sampling in order to construct a training set according to a desirable distribution over the classes. Essentially, text re-sampling can be viewed as providing new synthetic data that increase the training size of a class. Based on a corpus of newswire stories in English we present authorship identification experiments on various multi-class imbalanced cases.

1 INTRODUCTION

In recent years, researchers have paid increasing attention to authorship analysis in the framework of practical applications, such as verifying the authorship of emails and electronic messages, plagiarism detection in student essays, and forensic cases [1]. Authorship identification can be seen as a single-label multi-class text categorization problem where the candidate authors play the role of the classes. As concerns the text representation, various measures have been proposed in order to quantify the stylistic choices of the authors including function word frequencies, character *n*-gram frequencies, vocabulary richness measures, word-class frequencies, and syntactic analysis measures [2].

Very often, a common problem in authorship identification is the lack of sufficient text samples of undisputed authorship (at least for some of the candidate authors). On the other hand, a big amount of text samples may be available for other candidate authors. From a machine learning point of view, this constitutes the *class imbalance* problem (i.e., uneven distribution of the training set over the classes). This problem has been studied mainly within the framework of two-class datasets. The main approaches to deal with class imbalance attempt to re-balance the training set by performing either under-sampling of the majority class or over-sampling of the minority class. In general, the former approach has been proved to work better [3]. Alternatively, the sensitivity of the classification algorithm can be modified so that errors on minority class to be costlier than errors on majority class. Last but not least, the SMOTE approach creates new synthetic training data for the minority class [4]. This is achieved by adding a small value to some of the features of original training data and producing new data which lie close to the original ones in the multi-dimensional space of the problem.

Given a text categorization task, each training text is considered as a unit for constructing the training set. Usually, the length of the training texts is fixed or defined by the source of the documents. Moreover, text representations usually produce sparse data not quite suitable for a SMOTE-like approach. In this paper, we present methods to efficiently segment the training texts into sub-samples according to the size of the class. Textual data can be handled flexibly so that to produce a variable amount of text samples of variable length. Moreover, by using text re-sampling methods it is possible to provide new synthetic data.

2 AUTHORSHIP IDENTIFICATION

In this paper, we are based on the frequencies of character *n*-grams for text representation. Let $\mathbf{G}_d=\{g_1, g_2, \ldots, g_d\}$ be the ordered set (by decreasing frequency) of the most frequent character *n*-grams of the training set. Consider f_{ij} as the normalized frequency of the *j*-th *n*-gram of $\mathbf{G}_d$ in the *i*-th text. Then, a text x_i is represented as the vector $\langle f_{i1}, f_{i2}, \ldots, f_{id}\rangle$ (in this study, d=5,000 and n=3). A SVM, a model able to handle highly dimensional and sparse data, is then applied to these vectors. The Weka implementation was used with default parameters. Note that this approach is language-independent. However, for achieving best results given a particular corpus or natural language, one should explore the most appropriate amount and length of character *n*-grams.

An English corpus of newswire stories by 10 authors (100 texts per author) taken from the new Reuters Corpus Volume 1 has been used in this study. Apart from belonging to the same genre, all the texts fall under the CCAT topic (corporate/industrial) in order to minimize the factors that distinguish the classes despite authorship. Initially, this corpus was divided into non-overlapping training and test texts of equal size. In order to simulate the imbalance conditions of a real-world authorship identification case, a Gaussian distribution of training texts over the candidate authors is assumed. Given this setting, the *multi-class imbalance ratio* of the problem can be defined as *peak/base*, where *peak* and *base* are the size (in training texts) of the biggest and smallest classes, respectively. Figure 1 (left) shows an example distribution of the training set for *peak*=20 and *base*=5. Authors near the center of the distribution are the majority classes while the authors at both sides of the distribution are the minority classes. Different values of *base* and *peak* produce multi-class imbalanced datasets (see Table 1).

3 TESTED METHODS

When using all the training texts, the accuracy is 79.4% which is considered as the upper bound. On the other hand, when each training text is considered as unit and all training texts are used, a

[1] Dept. of Information and Communication Systems Eng., University of the Aegean, 83200, Karlovassi, Greece, email: stamatatos@aegean.gr

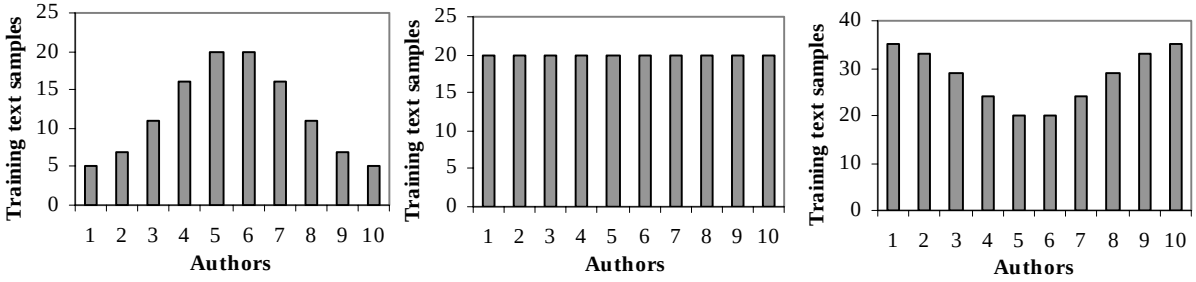

Figure 1. Distribution of text samples over the authors in multi-class training sets. Left: original imbalanced training text samples. Middle: Balanced training text samples produced by method-3. Right: Imbalanced training text samples produced by method-4.

Table 1. Accuracy results for the presented methods on different imbalanced cases together with upper (UB) and lower bounds (LB).

Case			Accuracy (%)					
Base	Peak	Ratio	UB	LB	M1	M2	M3	M4
2	10	5	79.4	51.2	52.4	56.8	56.6	**57.34**
2	20	10	79.4	47.4	52.4	**56.8**	51.8	55.38
2	50	25	79.4	53.6	52.4	56.8	54.6	**59.02**
5	10	2	79.4	60	49.8	59.8	**65.2**	58.52
5	20	4	79.4	56.4	49.8	59.8	59	**61.44**
5	50	10	79.4	59.6	49.8	59.8	62.4	**65.76**
10	20	2	79.4	62.8	61.4	**68.8**	65.6	66.18
10	50	5	79.4	65	61.4	68.8	68	**69.18**

lower bound of accuracy is provided for any imbalanced case. The following methods were tested (results in Table 1):

- Method-1: Under-sampling of the majority classes. For all authors, exactly *base* training texts were used.
- Method-2: Under-sampling of the majority classes using fixed length samples. All the available training texts were concatenated in one text per author. Let x_{min} be the size (in text lines) of the shortest author file. Then, the first x_{min} text lines of each author file were segmented into text sub-samples of length a (in text lines). It was observed that small values of a (2 or 3) tend to provide better results (in Table 1, a=3). Note that, in RCV1 each text line comprises one full sentence.
- Method-3: Re-balancing the dataset by variable length text samples. As previously, author files are produced. Let x_i and x_{max} be the text length (in text lines) of the i-th author and the longest author file, respectively. Then, each author file is segmented into text sub-samples of length x_{max}/x_i. Thus, a balanced dataset is produced having k text samples per class (in Table 1, k=50).
- Method-4: Re-balancing the dataset by text re-sampling. Again, author files are produced. A variable number of text samples per author is produced according to the length of their file. Many short text samples are produced for the minority classes and less but longer text samples are produced for the majority classes. The text lines included in a text sample are selected randomly and a text line may be included in more than one text sample (in Table 1, average accuracy after 50 runs is shown). Note that this method produces an imbalanced training set (in contrast to methods 1-3). However, the originally minority classes are now represented by more samples than the originally majority classes (see Figure 1).

Table 2 shows the identification accuracy per author (when *base*=5 and *peak*=20) as deviation from the baseline. The baseline method roughly resembles the distribution of training texts over the authors (the more training texts of one author, the better the results). Method-1 improves minority authors but fails to keep the majority authors on high level. Method-2 improves the minority authors without a dramatic loss in majority authors. Method-3 achieves to keep majority authors on high level (it even improves them) but it fails to significantly improve the minority authors. Finally, method-4 considerably improves minority authors with the cost of a slight reduction on the majority authors.

Table 2. The identification accuracy per author of the presented methods on imbalanced corpus (*base*=5, *peak*=20) expressed as deviation from the lower bound (LB).

Author	Tr. Set	LB	M1	M2	M3	M4
A01	5	36	+20	+36	+10	+18
A02	7	26	+22	+28	-16	+12
A03	11	66	-22	-30	-4	-18
A04	16	38	-16	+34	+6	-4
A05	20	100	-12	-8	0	0
A06	20	100	-18	-4	0	-8
A07	16	98	-74	-36	0	-10
A08	11	56	-26	-38	+12	-2
A09	7	6	+12	+10	0	+12
A10	5	38	+48	+42	+18	+46
Accuracy		56.4	-6.6	+3.4	+2.6	+4.6

4 CONCLUSION

Since textual data can be easily segmented in small pieces, they can be handled more flexibly in comparison to other kinds of data. Various text sampling and re-sampling methods were examined to handle multi-class imbalanced textual data. The main idea of the most successful method was to produce many short text samples for the minority classes and less but longer text samples for the majority classes. The basic methods presented here can be combined in order to further improve the results. For instance, method-3 and method-4 can be applied together. Alternatively, they could be used to train different classifiers which, then, can be combined in an ensemble scheme. Recall from Table 2 that the classification errors made by these methods are to a great extent uncorrelated, a crucial condition to build effective ensembles.

REFERENCES

[1] A. Abbasi and H. Chen, 'Applying Authorship Analysis to Extremist-Group Web Forum Messages' *IEEE Intelligent Systems*, **20**(5), 67-75, (2005).

[2] E. Stamatatos, N. Fakotakis, and G. Kokkinakis, 'Automatic Text Categorization in Terms of Genre and Author', *Computational Linguistics*, **26**(4), 471-495, (2000).

[3] N. Japkowicz, and S. Stephen, 'The Class Imbalance Problem: A Systematic Study', *Intelligent Data Analysis*, **6**(5), 429-450, (2002).

[4] N. Chawla, K. Bowyer, L. Hall, and W. Kegelmeyer, 'SMOTE: Synthetic Minority Over-sampling Technique'. *Journal of Artificial Intelligence Research*, **16**, 321-357, (2002).