Taking Advantage of the Web for Text Classification with Imbalanced Classes

Rafael Guzmán-Cabrera[1,2], Manuel Montes-y-Gómez[3],
Paolo Rosso[2], Luis Villaseñor-Pineda[3]

[1]FIMEE, Universidad de Guanajuato, Mexico
guzmanc@salamanca.ugto.mx
[2]DSIC, Universidad Politécnica de Valencia, Spain
prosso@dsic.upv.es
[3]LTL, Instituto Nacional de Astrofísica, Óptica y Electrónica, Mexico
{mmontesg, villasen}@inaoep.mx

Abstract. A problem of supervised approaches for text classification is that they commonly require high-quality training data to construct an accurate classifier. Unfortunately, in many real-world applications the training sets are extremely small and present imbalanced class distributions. In order to confront these problems, this paper proposes a novel approach for text classification that combines under-sampling with a semi-supervised learning method. In particular, the proposed semi-supervised method is specially suited to work with very few training examples and considers the automatic extraction of untagged data from the Web. Experimental results on a subset of Reuters-21578 text collection indicate that the proposed approach can be a practical solution for dealing with the class-imbalance problem, since it allows achieving very good results using very small training sets.

1 Introduction

Nowadays there is a lot of digital information available from the Web. This situation has produced a growing need for tools that help people to find, filter and analyze all these resources. In particular, text classification [10], the assignment of free text documents to one or more predefined categories based on their content, has emerged as a very important component in many information management tasks.

The state-of-the-art approach for automatic text classification considers the application of a number of statistical and machine learning techniques, including regression models, Bayesian classifiers, support vector machines, nearest neighbor classifiers, neuronal networks and statistical methods driven by a hierarchical topic dictionary to mention some [1, 10, 3]. A major difficulty with this kind of supervised techniques is that they commonly require high-quality training data to construct an accurate classifier. Unfortunately, in many real-world applications the training sets are *extremely small* and even worst, they present *imbalanced class distributions* (i.e., the

This work was done under partial support of CONACYT-Mexico (43990) MCyT-Spain (TIN2006-15265-C06-04) and PROMEP (UGTO-121).

number of examples in some classes are significantly greater that the number of examples in the others).

In order to overcome these problems, recently many researches have been working on semi-supervised learning algorithms as well as on different solutions to the class-imbalance problem (for an overview refer to [2, 11]). On the one hand, it has been showed that by augmenting the training set with additional –unlabeled– information it is possible to improve the classification accuracy using different learning algorithms such as naïve Bayes [9], support vector machines [7], and nearest-neighbor algorithms [13]. On the other hand, it has also been demonstrated that by adjusting the number of examples in the majority or minority classes it is possible to tackle the suboptimal classification performance caused by the class-imbalance [6]. In particular, there is evidence that under-sampling, a method in which examples of the majority classes are removed, leads to better results than over-sampling, a method in which examples from the minority classes are duplicated [5].

In this paper we propose a novel approach for text classification with imbalanced classes that combines under-sampling and semi-supervised methods. The idea is to use under-sampling to balance an original imbalanced training set, and then apply a semi-supervised classification method to compensate the missing of information by adding new –highly discriminative– training instances.

The most relevant component of the proposed approach is the semi-supervised classification method. It mainly differs from previous methods in three main concerns. First, it is specially suited to work with *very* few training examples. Whereas previous methods consider hundreds of training examples, our method allows working with just groups of ten labeled examples per class. Second, it does not require a predefined set of unlabeled examples. It considers the automatic extraction of related untagged data from the Web. Finally, given that it deals with few training examples, it does not aim including a lot of additional information in the training phase; instead, it only incorporates a small group of examples that considerably augment the dissimilarities among classes.

It is important to mention that this method achieved very good results on classifying news documents about natural disasters [4]. It could construct an accurate classifier starting from only ten training examples per class. However, in that case, the training collection was simple: it only contained five clearly separable classes with no imbalance. In contrast, in this new experiment we aim to explore the capacity of the method to deal with more complex document collections that contain a great number of imbalanced and overlapped classes.

The rest of the paper is organized as follows. Section 2 shows the general scheme of the proposed approach for text classification with imbalanced classes. Section 3 describes the Web-based semi-supervised classification method. Section 4 presents some evaluation results on a subset of Reuters-21578 text collection. Finally, section 5 depicts our conclusions and future work.

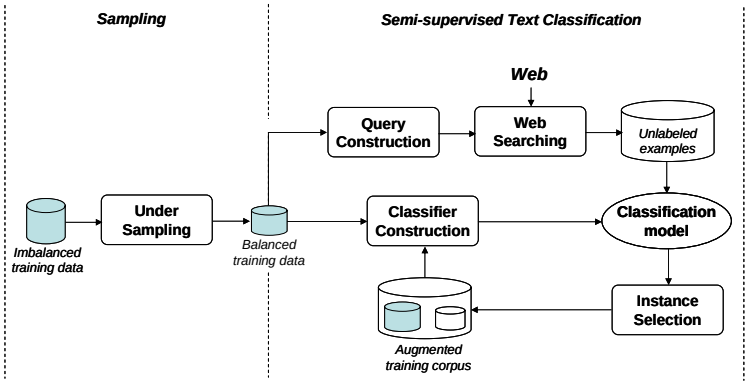

Figure 1. General overview of the approach

2 Overview of the Proposed Approach

Figure 1 shows the general scheme of the proposed approach. It consists of two main phases: under-sampling and semi-supervised text classification.

Under-sampling is one of the methods most commonly used to adapt machine-learning algorithms to imbalanced classes. As we mentioned, it considers the elimination of training examples from the majority classes. In this case, examples to be removed can be randomly selected, or near miss examples, or examples that are far from the minority of the class instances [5].

In our particular case, we apply a kind of "extreme" under-sampling over the original data set. The idea is to assemble a *small* balanced training corpus by eliminating –at random– a great number of examples from all classes. This extreme strategy was mainly motivated by the fact that small training sets are more advantageous for our semi-supervised classification method. In addition, this decision was also motivated by our interest on demonstrating that the problem of learning from imbalanced classes can be modeled as one of learning from *very* small training sets.

The second phase considers the semi-supervised classification method. This method consists of two main processes. The first one deals with the corpora acquisition from the Web, while the second one focuses on the semi-supervised learning problem. The following section describes in detail these two processes.

It is important to point out that the Web has been lately used as a corpus in many natural language tasks [8]. In particular, Zelikovitz and Kogan [14] proposed a method for mining the Web to improve text classification by creating a background text set. Our proposal is similar to this approach in the sense of it also mines the Web for additional information (extra-unlabeled examples). Nevertheless, our method applies finer procedures to construct the set of queries related to each class and to combine the downloaded information.

3 Semi-supervised Text Classification

3.1 Corpora Acquisition

This process considers the automatic extraction of unlabeled examples from the Web. In order to do this, it first constructs a number of queries by combining the most significant words for each class; then, using these queries it looks at the Web for some additional training examples related to the given classes.

Query Construction. In order to form queries for searching the Web, it is necessary to previously determine the set of relevant words for each class in the training corpus. The criterion used for this purpose is based on a combination of the frequency of occurrence and the information gain of words. We consider that a word w_i is relevant for class C if:

1. The frequency of occurrence of w_i in C is greater than the average occurrence of all words (happening more than once) in that class. That is:

$$f_{w_i}^{C} > \frac{1}{|C'|} \sum_{\forall w \in C'} f_w^C \text{ , where } C' = \left\{ w \in C \middle| f_w^C > 1 \right\}$$

2. The information gain of w_i with respect to C is positive. That is, if $IG_{w_i}^{C} > 0$.

Once obtained the set of relevant words per class, it is possible to construct the corresponding set of queries. Founded on the method by Zelikovitz and Kogan [14], we decide to construct queries of three words. This way, we create as many queries per class as all three-word combinations of its relevant words. We measure the significance of a query $q = \{w_1, w_2, w_3\}$ to the class C as indicated below:

$$\Gamma_C(q) = \sum_{i=1}^{3} f_{w_i}^{C} \times IG_{w_i}^{C}$$

Web Searching. The next action is using the defined queries to extract from the Web a set of additional unlabeled text examples. Based on the observation that most significant queries tend to retrieve the most relevant web pages, our method for searching the Web determines the number of downloaded examples per query in a direct proportion to its Γ-value. Therefore, given a set of M queries $\{q_1,\ldots, q_M\}$ for class C, and considering that we want to download a total of N additional examples per class, the number of examples to be extracted by a query q_i is determined as follows:

$$\Psi_C(q_i) = \frac{N}{\sum_{k=1}^{M} \Gamma_C(q_k)} \times \Gamma_C(q_i)$$

3.2 Semi-supervised learning

As we previously mentioned, the purpose of this process is to increase the classification accuracy by gradually augmenting the originally small training set with the ex-

amples downloaded from the Web. Our algorithm for semi-supervised learning is an adaptation of a method proposed elsewhere [12]. It mainly considers the following steps:

1. Build a weak classifier (C_l) using a specified learning method (l) and the training set available (T).
2. Classify the downloaded examples (E) using the constructed classifier (C_l). In order words, estimate the class for all downloaded examples.
3. Select the best m examples ($E_m \subseteq E$) based on the following two conditions:
    a. The estimate class of the example corresponds to the class of the query used to download it. In some way, this filter works as an ensemble of two classifiers: C_l and the Web (expressed by the set of queries).
    b. The example has one of the m-highest confidence predictions.
4. Combine the selected examples with the original training set ($T \leftarrow T \cup E_m$) in order to form a new training set. At the same time, eliminate these examples from the set of downloaded instances ($E \leftarrow E - E_m$).
5. Iterate σ times over steps 1 to 4 or repeat until $E_m = \varnothing$. In this case σ is a user specified threshold.
6. Construct the final classifier using the enriched training set.

4 Experimental Evaluation

4.1 Experimental Setup

Corpus. We selected the subset of the 10 largest categories of the Reuters-21578 corpus. In particular, we considered the ModApte split distribution, which includes all labeled documents published before 04/07/87 as training data (i.e., 7206 documents) and all labelled documents published after 04/07/87 as testing set (i.e., 3220 documents). Table 1 shows some numbers on this collection.

Table 1. Training/testing data sets

Category	Training Set	Test Set
ACQ	1650	798
CORN	182	71
CRUDE	391	243
EARN	2877	1110
GRAIN	434	194
INTEREST	354	159
MONEY-FX	539	262
SHIP	198	107
TRADE	369	182
WHEAT	212	94
Total	***7206***	***3220***

Searching the Web. We used Google as search engine. We downloaded 2,400 additional examples (snippets for these experiments) per class.

Learning method. We selected naïve Bayes (NB) as the base classification method.

Document Preprocessing. We removed all punctuation marks and numerical symbols, that is, we only considered alphabetic tokens. We also removed stop words and hapax legomena, and converted all tokens to lowercase. On the other hand, in all experiments we took the 1000 most frequent words as classification features.

Evaluation measure. The effectiveness of the method was measured by the classification accuracy, which indicates the percentage of documents that have been correctly classified from the entire document set.

Baseline. For this case, all training data was used to construct a naïve Bayes classifier. The achieved accuracy of this classifier over the given test data was of *84.7%*.

4.2 Experimental Results

As we mentioned in section 2, the proposed approach has two main phases: under-sampling and semi-supervised text classification. The idea is to apply under-sampling to assemble a balanced training corpus, and then use a semi-supervised classification method to compensate the missing of information by adding new –highly discriminative– training instances (i.e., snippets downloaded from the web).

Because our semi-supervised method is specially suited to work with *very few* training examples, we applied an "extreme" under-sampling over the original training data. Table 2 shows the accuracy results corresponding to different levels of data reduction. It is important to notice that using only 100 training examples per class it was practically possible to reach the baseline result.

Table 2. Accuracy percentage for different training data sets

Training examples per class	Accuracy percentage
10	58.6
20	73.7
30	77.3
40	79.3
50	81.8
80	82.8
100	84.1
182	84.0
Baseline	*84.7*

In order to evaluate the semi-supervised classification method we performed *two experiments*. The first one only used 10 training examples per class, whereas the other one employed 100 training instances per class.

It is important to clarify that using more examples allows constructing more general and consequently more relevant queries. For instance, using one hundred examples about the INTEREST class, we constructed queries such as: <*bank* + *money* + *interest*>, <*money* + *market* + *banks*> and <*bank* + *interest* + *rate*>.

Using the automatically constructed queries, we collected from the Web a set of 2,400 snippets per class, obtaining a total of 24,000 additional unlabeled examples. It is interesting to point out that thanks to the snippet's small size (that only considers

the immediate context of the query's words), the additional examples tend to be lees ambiguous and contain several valuable words that are highly related with the topic at hand. As an example, look at the following snippet for the class *INTEREST*:

> *<compare mortgage rates home loans cd rates auto loans credit free objective information rate quotes consumer bank products cds auto loans home equity loans money market funds personal loans>*

Finally, the downloaded snippets were classified using the original document collection as training set (refer to section 3.2). The best *ten* examples per class, i.e., those with more confidence predictions, were selected at each iteration and were incorporated to the original training set in order to form a new training collection. In both experiments, we performed 10 iterations. Table 3 shows the accuracy results for all iterations of both experiments.

Table 3. Accuracy percentage after the training corpus enrichment

Labeled Training Instances	Base Accuracy	Iteration									
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
10	58.6	66.9	68.7	69.6	70.3	**70.6**	68.6	69.0	69.0	68.5	68.7
100	84.1	84.6	84.7	84.8	86.6	86.8	86.8	**86.9**	86.7	86.7	86.7

From table 3 we can observe that the semi-supervised learning method did its job. For instance, when using only 10 training examples per class the method produced a notable 12% increase in the accuracy (from 56.6 to 70.6). Nevertheless, it is clear that given the complexity of the given test collection (that contains some semantically related classes such as grain, corn and wheat) it is necessary to start with more training examples.

In the case of the second experiment (which made use of 100 training examples per class), the increment in the accuracy was not as high as in the first experiment. It only moved the accuracy from 84% to 86.9%. However, it is important to mention that this increment was enough to outperform the baseline result. In other words, the method allowed obtaining a better accuracy using only 1000 training examples than considering all 7206.

5 Conclusions and Future Work

This paper proposed a novel approach for text classification with imbalanced classes that combines under-sampling and semi-supervised learning methods. The general idea of the approach is to use under-sampling to balance an original imbalanced training set, and then apply a semi-supervised classification method to compensate the missing of information by adding new –highly discriminative– training instances.

In particular, the most relevant component of the approach is the semi-supervised text classification method. This method differs from others in that: (*i*) it is specially suited to work with *very* few training examples, (*ii*) it automatically collects from the Web the unlabeled data and, (*iii*) it only incorporates into the training phase a small group of highly discriminative unlabeled examples.

In general, the achieved results allow us to formulate the following conclusions. On the one hand, the proposed combined approach can be a practical solution for the problem of text classification with imbalanced classes. On the other hand, our Web-based semi-supervised learning method is a quite pertinent tool for text classification, since it allows achieving very good results using very small training sets.

As future work we plan to apply the proposed approach to other collections with higher imbalance rates, for instance, to a different subset of the Reuters corpus. Also, given that the highest accuracies were obtained before completing all possible iterations, we aim to study the behavior of the iterative semi-supervised learning process in order to define a better stop criterion. Finally, we also plan to evaluate the approach, in particular, the semi-supervised learning method, in some non-topical classification problems such as authorship attribution and genre detection.

References

1. Aas K., and Eikvil L., Text Categorization: A survey, Technical Report, number 941, Norwegian Computing Center, 1999.
2. Chawla N. V., Japkowicz N., Kolcz A., Editorial: Special Issue on Learning from Imbalanced data Sets. ACM SIGKDD Exploration Newsletters. Volume 6, Issue 1, June 2004.
3. Gelbukh A., Sidorov G., Guzman-Arénas A., Use of a Weighted Topic Hierarchy for Document Classification. Lecture Notes in Artificial Intelligence, No. 1692, 1999, Springer, pp. 130-135.
4. Guzmán-Cabrera R., Montes-y-Gómez M., Rosso P., Villaseñor-Pineda L., Improving Text Classification by Web Corpora. Advances in Soft Computing, No. 43, 2007, Springer, pp. 154-159.
5. Hoste V., Optimization Issues in Machine Learning of Coreference Resolution. Doctoral Thesis, Faculteit Letteren en Wijsbegeerte, Universiteit Antwerpen, 2005.
6. Japkowicz N., Learning from Imbalanced Data Sets: A comparison of Various Strategies. AAAI Workshop on Learning from Imbalanced Data Sets. Tech Rep. WS-00-05, Menlo Park, CA: AAAI Press, 2000.
7. Joachims T., Transductive inference for text classification using support vector machines, Proceedings of the Sixteenth International Conference on Machine Learning, 1999.
8. Kilgarriff A., and Greffenstette G., Introduction to the Special Issue on Web as Corpus, Computational Linguistics, 29(3), 2003.
9. Nigam K., Mccallum A. K., Thrun S., and Mitchell T., Text classification from labeled and unlabeled documents using EM, Machine Learning, 39(2/3):103–134, 2000.
10. Sebastiani F., Machine learning in automated text categorization, ACM Computing Surveys, 34(1):1–47, 2002.
11. Seeger M., Learning with labeled and unlabeled data. Technical report, Institute for Adaptive and Neural Computation, University of Edinburgh, Edinburgh, United Kingdom, 2001.
12. Solorio T., Using unlabeled data to improve classifier accuracy, Master Degree Thesis, Computer Science Department, INAOE, Mexico, 2002.
13. Zelikovitz S., and Hirsh H., Integrating background knowledge into nearest-Neighbor text classification, In Advances in Case-Based Reasoning, ECCBR Proceedings, 2002.
14. Zelikovitz S., and Kogan M., Using Web Searches on Important Words to Create Background Sets for LSI Classification, 19th International FLAIRS conference, Melbourne Beach, Florida, May 2006.