Improving Text Classification Performance with Incremental Background Knowledge

Catarina Silva[1,2] and Bernardete Ribeiro[2]

[1] School of Technology and Management,
Polytechnic Institute of Leiria, Portugal
[2] Dep. Informatics Eng., Center Informatics and Systems,
Univ. of Coimbra, Portugal
catarina@dei.uc.pt, bribeiro@dei.uc.pt

Abstract. Text classification is generally the process of extracting interesting and non-trivial information and knowledge from text. One of the main problems with text classification systems is the lack of labeled data, as well as the cost of labeling unlabeled data. Thus, there is a growing interest in exploring the use of unlabeled data as a way to improve classification performance in text classification. The ready availability of this kind of data in most applications makes it an appealing source of information.

In this work we propose an Incremental Background Knowledge (IBK) technique to introduce unlabeled data into the training set by expanding it using initial classifiers to deliver oracle decisions. The defined incremental SVM margin-based method was tested in the Reuters-21578 benchmark showing promising results.

1 Introduction

Applications of text mining are ubiquitous, since almost 80% of the information available is stored as text. Thus, there is an effective interest in researching and developing applications that better help people handling text-based information. On the other hand, the wealth of text in digital form has made the organization of that information into a complex and vitally important task.

Most text categorization methods, e.g., K-Nearest Neighbor, Naïve Bayes, Neural Nets and Support Vector Machines, have their performance greatly defined by the training set available. To achieve the best classification performance with a machine learning technique, there has to be enough labeled data. However, these data are costly and sometimes difficult to gather. This is one key difficulty with current text categorization algorithms, since they require manual labeling of more documents than a typical user can tolerate [1].

Labeling data is expensive but, in most text categorization tasks, unlabeled data are often inexpensive, abundant and readily available. Therefore, to achieve the purpose of using relatively small training sets, the information that can be extracted from the testing set, or even unlabeled examples, is being investigated as a way to improve classification performance [2,3]. Seeger in [4] presents a report on learning with unlabeled data that compares several approaches.

C. Alippi et al. (Eds.): ICANN 2009, Part I, LNCS 5768, pp. 923–931, 2009.

In general, unlabeled examples are much less expensive and easier to gather than labeled ones. This is particularly true for text classification tasks involving online data sources, such as web pages, email and news stories, where large amounts of text are readily available. Collecting this text can frequently be done automatically, so it is feasible to collect a large set of unlabeled examples. If unlabeled examples can be integrated into supervised learning, then building text classification systems will be significantly faster, less expensive and more effective.

There is a catch however, because, at first glance, it might seem that nothing is to be gained from unlabeled data, since an unlabeled document does not contain the most important piece of information - its classification.

Consider the following example to give some insight of how unlabeled data can be useful. Suppose we are interested in recognizing web pages about conferences. We are given just a few conferences and non-conferences web pages, along with a large number of pages that are unlabeled. By looking at just the labeled data, we determine that pages containing the word *paper* tend to be about conferences. If we use this fact to estimate the classification of the many unlabeled web pages, we might find that the word *deadline* occurs frequently in the documents that are classified in the positive class. This co-occurrence of the words *paper* and *deadline* over the large set of unlabeled training data can provide useful information to construct a more accurate classifier that considers both *paper* and *deadline* as indicators of positive examples.

In this work we propose an Incremental Background Knowledge (IBK) technique that uses the Support Vector Machine (SVM) classification margin to determine unlabeled examples classification and strengthen the training set. The IBK is an improvement of a Basic Background Knowledge (BBK) approach already proposed by the authors in [5]. The new incremental technique provides a more stable convergence to an improved performance without using any new supervisor knowledge.

The rest of the paper is organized as follows. Section 2 addresses several text classification issues, setting guidelines for problem formulation, including the application of Support Vector Machines (SVMs) to text classification tasks.

Section 3 presents the proposed Incremental Background Knowledge (IBK) approach comparing with the previous technique, followed by experimental setup and results in Sections 4 and 5 respectively. Finally, Section 6 presents some conclusions and future work.

2 Text Classification

The goal of text classification is the automatic assignment of documents to a fixed number of semantic categories. Each document can be in multiple, exactly one, or no category at all. Using machine learning, the objective is to learn classifiers from examples, which assign categories automatically. This is usually considered a supervised learning problem. To facilitate effective and efficient learning, each category is treated as a separate binary classification problem. Each of such problems answers the question of whether or not a document should be assigned to a particular category [6].

Documents, which typically are strings of characters, have to be transformed into a suitable representation both for the learning algorithm and the classification task. The most common representation is known as the *Bag of Words* and represents a document by the words occurring in it. Usually the irrelevant words are filtered using a stopword list and the word ordering is not deemed relevant for most applications. Information retrieval investigation proposes that instead of words, the units of representation could be word stems. A word stem is derived from the occurrence form of a word by removing case and inflection information. For example "viewer", "viewing", and "preview" are all mapped to the same stem "view".

This leads to an attribute-value representation of text. Each distinct word w_i corresponds to a feature $TF(w_i, x)$, representing the number of times word w_i occurs in the document x. Refining this basic representation, it has been shown that scaling the dimensions of the feature vector with their inverse document frequency $IDF(w_i)$ leads to an improved performance. $IDF(w_i)$ (1) can be calculated from the document frequency $DF(w_i)$, which is the number of documents the word w_i occurs in.

$$IDF(w_i) = log\left(\frac{D}{DF(w_i)}\right) \tag{1}$$

Here, D is the total number of documents. The inverse document frequency of a word is low if it occurs in many documents and is highest if the word occurs in only one. To disregard different document lengths, each document feature vector $\mathbf{x}$ is normalized to unit length [7].

2.1 SVM Text Classification

Support Vector Machines (SVMs) are a learning method introduced by Vapnik [8] based on his Statistical Learning Theory and Structural Risk Minimization Principle. When using SVMs for classification, the basic idea is to find the optimal separating hyperplane between the positive and negative examples. The optimal hyperplane is defined as the one giving the maximum margin between the training examples that are closest to it. Support vectors are the examples that lie closest to the separating hyperplane. Once this hyperplane is found, new examples can be classified simply by determining on which side of the hyperplane they are.

Although text categorization is a multi-class, multi-label problem, it can be broken into a number of binary class problems without loss of generality. This means that instead of classifying each document into all available categories, for each pair $\{document, category\}$ we have a two class problem: the document either belongs or does not to the category. Although there are several linear classifiers that can separate both classes, only one, the Optimal Separating Hyperplane, maximizes the margin, i.e., the distance to the nearest data point of each class, thus presenting better generalization potential.

The output of a linear SVM is $u = \mathbf{w} \times \mathbf{x} - b$, where $\mathbf{w}$ is the normal weight vector to the hyperplane and $\mathbf{x}$ is the input vector. Maximizing the margin can be seen as an optimization problem:

$$\begin{aligned} &minimize \quad \frac{1}{2}||\mathbf{w}||^2, \\ &subjected\ to\ \ y_i(\mathbf{w}.\mathbf{x}+b) \geq 1, \forall i, \end{aligned} \tag{2}$$

where $\mathbf{x}$ is the training example and y_i is the correct output for the ith training example. Intuitively the classifier with the largest margin will give low expected risk, and hence better generalization.

3 Incremental Background Knowledge

Some authors [9] refer to unlabeled data as background knowledge, defining it as any unlabeled collection of text from any source that is related to the classification task. Joachims presents in [6] a study on transductive SVMs (TSVMs) introduced by Vapnik [8]. TSVMs make use of the testing set and extend inductive SVMs, finding an optimal separating hyperplane not only of the training examples, but also of the testing examples [10].

The Incremental Background Knowledge (IBK) technique we now propose is in fact a development of a Basic Background Knowledge (BBK) approach already proposed by the authors in [5]. We will start by generally describing the basic strategy that will serve as base comparison, and then the incremental approach that constitutes the main contribution of this work (more details on BBK can be found in [5]).

In the BBK, first an inductive SVM classifier (see Section 2.1) is inferred from the training set, and then it is applied to the unlabeled examples. The BBK approach incorporates, in the training set, new examples classified by the SVM with larger margin, which can be assumed as the ones where the SVM classifier presents more confidence. Fig. 1 illustrates an example where four unlabeled examples (black dots) are classified with small and large margins.

Formally, the BBK approach proceeds by incorporating unlabeled examples (only the features, not the classification) from the unlabeled/testing set directly

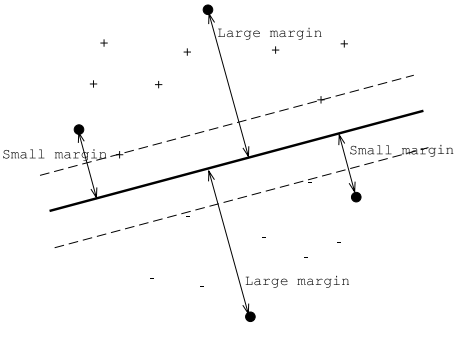

Fig. 1. Unlabeled examples (black dots) with small and large margins

into the training set as classified by the baseline inductive SVM, i.e., an example $(\mathbf{x}_i, y_i)$ will be chosen if Equation (3) holds:

$$(\mathbf{x}_i, y_i) : \rho(\mathbf{x}_i, y_i) = \frac{2}{\|w\|} > \Delta, \tag{3}$$

Δ was heuristically defined. Notice that Δ is intrinsically related to the margin, i.e when Δ is decreased, in fact we are decreasing the classification margin of accepted unlabeled examples and thus accepting examples classified with less confidence. This level of confidence should depend on the capabilities of the base classifier, or in other words, the better the base classifier the lower we can set the threshold on Δ (and thus on the margin) to introduce newly classified unlabeled examples into the training set.

In the IBK approach we now suggest, a structural change is proposed to deal with the weaker point of the BBK technique, i.e. the definition of Δ. We proposed the iterative procedure illustrated in Fig. 2. As can be gleaned from this figure, the training set is incrementally constructed by iteratively decreasing the value of Δ, i.e. reducing the confidence threshold for an unlabeled example to be added as classified by the SVM. This approach rational is that as Δ is decreased, the classifiers are also getting better due to the additional information in the training set, thus justifying lowering the confidence threshold. Algorithm 1 below more formally defines the IBK procedure.

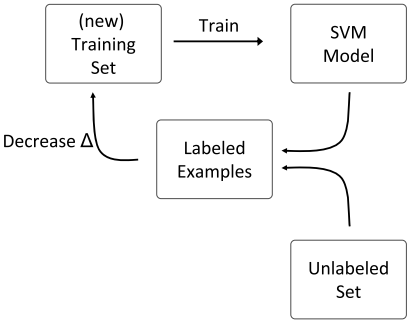

Fig. 2. Proposed approach: Incremental Background Knowledge

Algorithm 1. Incremental Background Knowledge Algorithm

Current training set ← Initial dataset
Δ ← initial Δ value
WHILE not all unlabeled examples added
    Infer an SVM classifier with current training set
    Classify unlabeled examples with the classifier
    Select the newly classified examples with margin larger that Δ
    Add the selected examples to the current training set
    Decrease Δ
ENDWHILE

4 Experimental Setup

4.1 Reuters-21578 Benchmark

The widely accepted Reuters-21578 benchmark was used in the experiments. It is a financial corpus with news articles documents averaging 200 words each. Reuters-21578 is publicly available at http://kdd.ics.uci.edu/databases/reuters-21578/reuters21578.html. In this corpus 21,578 documents are classified in 118 categories.

Reuters is a very heterogeneous corpus, since the number of documents assigned to each category is very variable. There are documents not assigned to any of the categories and documents assigned to more than 10 categories. On the other hand, the number of documents assigned to each category is also not constant. There are categories with only one assigned document and others with thousands of assigned documents. The ModApte split was used, using 75% of the articles (9603 items) for training and 25% (3299 items) for testing. Table 1 presents the 10 most frequent categories and the number of positive training and testing examples. These 10 categories are widely accepted as a benchmark, since 75% of the documents belong to at least one of them.

Table 1. Number of positive training and testing documents for the Reuters-21578 most frequent categories

Category	Train	Test
Earn	2715	1044
Acquisitions	1547	680
Money-fx	496	161
Grain	395	138
Crude	358	176
Trade	346	113
Interest	313	121
Ship	186	89
Wheat	194	66
Corn	164	52

4.2 Performance Metrics

In order to evaluate a binary decision task we first define a contingency matrix representing the possible outcomes of the classification, as shown in Table 2.

Several measures have been defined based on this contingency table, such as, error rate ($\frac{b+c}{a+b+c+d}$), recall ($\frac{a}{a+c}$), and precision ($\frac{a}{a+b}$), as well as combined measures, such as, the van Rijsbergen F_β measure [11], which combines recall and precision in a single score, $F_\beta = \frac{(\beta^2+1)P \times R}{\beta^2 P + R}$. The latter is one of the best

Table 2. Contingency table for binary classification

	Class Positive	Class Negative
Assigned Positive	a	b
	(True Positives)	(False Positives)
Assigned Negative	c	d
	(False Negatives)	(True Negatives)

suited measures for text classification used with $\beta = 1$, i.e. F_1, and thus the results reported in this paper are macro-averaged F_1 values.

5 Experimental Results

Experiments were carried out varying the values of Δ starting with no inclusion of new unlabeled examples until practically all available examples were added. In the BBK approach, for each value of Δ a new training set was constructed, learned and tested, while for IBK the training set in each iteration, corresponding to a value of Δ, was used as baseline for the next iteration, where it was again incremented (see Fig. 2).

Fig. 3 shows the F1 performance for both approaches and for the several values of Δ. Notice that the values of Δ in x-axis are decreasing values, reflecting the nature of the IBK technique that starts to add large margin classified examples (with high confidence and large Δ values) and proceeds with decreasing values of margin, confidence and Δ. It is clear from this figure that the IBK generally surpasses BBK. However, the most compelling analysis is in its stability

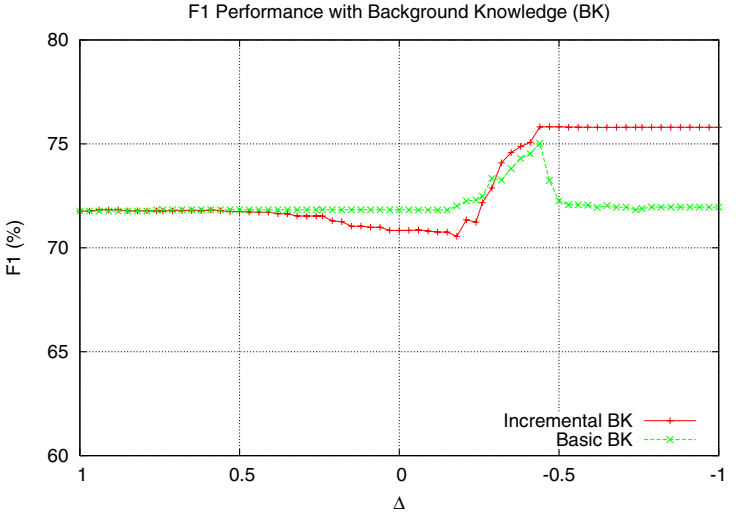

Fig. 3. F1 performance for IBK and BBK for different Δ values

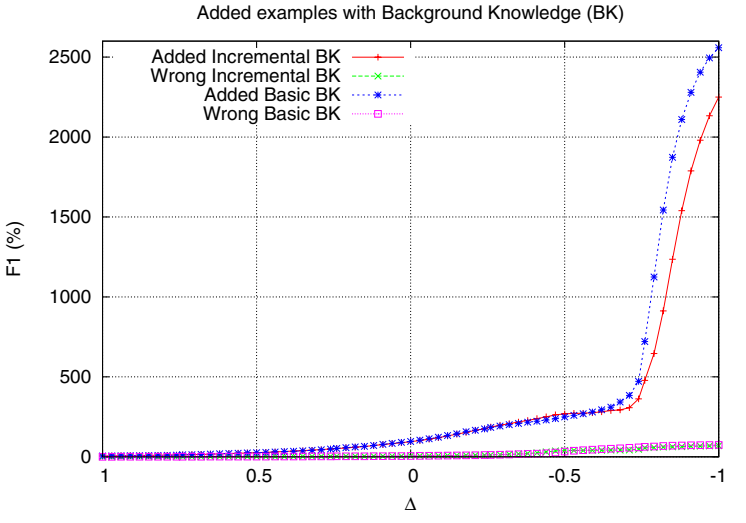

Fig. 4. Total added examples and wrongly added examples for IBK and BBK for different Δ values

to different values of confidence. While BBK presents sensitivity when the confidence drops below an acceptable threshold, IBK remains fairly insensitive and stable, even when all examples are added. It is possible to assert the number of added examples and of these how many are wrongly classified in the graphic in Fig. 4. It is interesting to notice that although more examples are added in the BBK, the number of examples introduced with the wrong classification in the training sets are fairly low and equivalent, despite the difference in classification performance.

6 Conclusions and Future Work

In this work we proposed an Incremental Background Knowledge (IBK) technique that uses the Support Vector Machine (SVM) classification margin to determine unlabeled examples classification and strengthen the training set. The IBK is an improvement of a Basic Background Knowledge (BBK) approach previously developed by the authors.

The main contribution is in the area of semi-supervised learning, by devising a stable and efficient mechanism for automatically incorporating unlabeled examples in the learning task. In fact, the incremental approach does not limit the number of examples introduced and requires no feedback from an oracle. Results shown that the improvement obtained in an optimal point of operation is at least maintained even in severe circumstances of very low level of confidence in the added unlabeled examples.

Future work is expected in further validating the strategy in different applications, namely multiclass applications.

References

1. Schohn, G., Cohn, D.: Less is more: Active Learning with Support Vector Machines. In: International Conference on Machine Learning, pp. 839–846 (2000)
2. Hong, J., Cho, S.-B.: Incremental Support Vector Machine for Unlabeled Data Classification. In: International Conference on Neural Information Processing (ICONIP), pp. 1403–1407 (2002)
3. Liu, B., Dai, Y., Li, X., Lee, W., Yu, P.: Building Text Classifiers Using Positive and Unlabeled Examples. In: International Conference on Data Mining, pp. 179–188 (2003)
4. Seeger, M.: Learning with Labeled and Unlabeled Data, Technical Report, Institute for Adaptive and Neural Computation, University of Edinburgh (2001)
5. Silva, C., Ribeiro, B.: On Text-based Mining with Active Learning and Background Knowledge using SVM. Journal of Soft Computing - A Fusion of Foundations, Methodologies and Applications 11(6), 519–530 (2007)
6. Joachims, T.: Transductive Inference for Text Classification using Support Vector Machines. In: International Conference on Machine Learning, pp. 200–209 (1999)
7. Sebastiani, F.: A Tutorial on Automated Text categorisation. In: Amandi, A., Zunino, A. (eds.) Proceedings of ASAI 1999, 1st Argentinian Symposium on Artificial Intelligence, Buenos Aires, AR, pp. 7–35 (1999)
8. Vapnik, V.: The Nature of Statistical Learning Theory, 2nd edn. Springer, Heidelberg (1999)
9. Zelikovitz, S., Hirsh, H.: Using LSI for text classification in the presence of background text. In: Tenth International Conference on Information Knowledge Management, pp. 113–118 (2001)
10. Silva, C., Ribeiro, B.: Labeled and Unlabeled Data in Text Categorization. In: IEEE International Joint Conference on Neural Networks (2004)
11. van Rijsbergen, C.: Information Retrieval, 2nd edn. Butterworths, London (1979)