Exploring a New Space of Features for Document Classification: Figure Clustering

Nawei Chen, Hagit Shatkay and Dorothea Blostein

School of Computing, Queen's University, Kingston, Ontario, Canada

Abstract

Automatic document classification is an important step in organizing and mining documents. Information in documents is often conveyed using both text and images that complement each other. Typically, only the text content forms the basis for features that are used in document classification. In this paper, we explore the use of information from figure images to assist in this task. We explore image clustering as a basis for constructing *visual words* for representing documents. Once such visual words are formed, the standard bag-of-words representation along with commonly used classifiers, such as the naïve Bayes, can be used to classify a document. We report here results from classifying biomedical documents that were previously used in the TREC Genomics track, employing the image-based representation. Efforts are ongoing to improve image-based classification and analyze the relationships between text and images. The goal is to develop a new set of features to supplement current text-based features.

1 Introduction

Automatic document classification is an important step in organizing documents and in literature mining. The current growth of digital libraries along with increase in the number of web publications, leads to much research in this area. However, documents convey information using not only text – but also image data. These two modalities typically complement each other. This point is well-illustrated in the thumbnail of a document from a biomedical journal, shown in Figure 1, in which both images and text are used to produce a complete report. While text classification is a mature field [3], image-based document classification is relatively unexplored. In this paper, we investigate the use of image features extracted from figures and illustrations for document classification.

Figure 1. Thumbnail of an example document, which conveys information using both text and images.

Our experiments focus on biomedical document classification, which is central for supporting curation tasks in biological databases, such as those of the Mouse Genome Institute (MGI)[1]. We make use of the dataset of full-text documents provided by the TREC Genomics track 2005 [2]. During the years 2004 and 2005, the TREC Genomics track defined challenges that simulated some of the tasks performed by MGI curators. Training and test datasets labeled by MGI human experts were provided, along with objective evaluation metrics [2]. Documents usually contain several figures and illustrations. A figure may consist of a few meaningful subfigures. We propose and explore here a method to represent

[1] The MGI (Mouse Genome Informatics) system is an initiative of the Jackson Labs (http://www.informatics.jax.org/). It provides integrated access to data on the genetics, genomics and biology of the laboratory mouse.

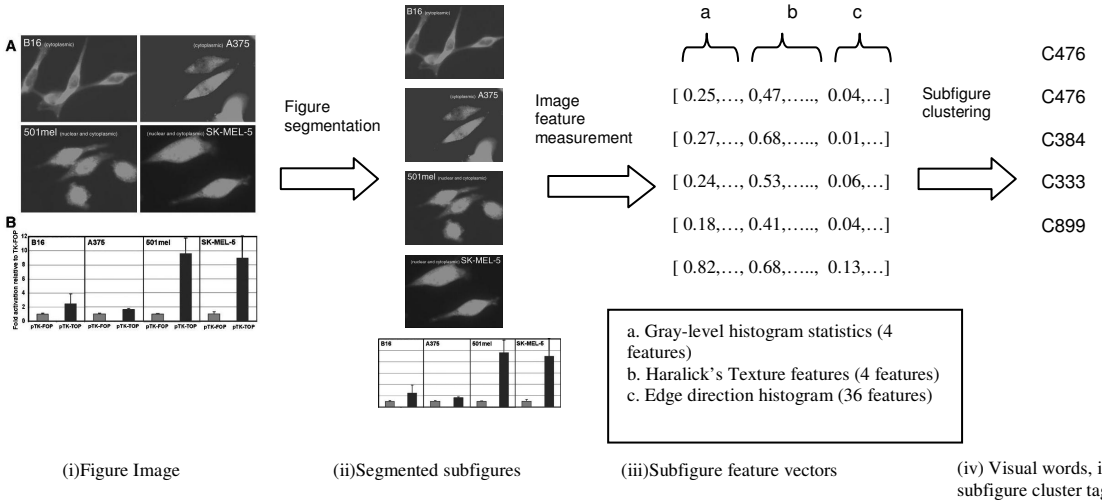

Figure 2. (i) An example figure extracted from the document shown in Figure 1 (PubMed Identifier 12235125 [5]. Figures reproduced with permission of the Rockefeller University press). (ii) Figure segmentation results. Connected components whose bounding box area is too small are discarded since they are most likely characters used to label figures. (iii) Low-level image feature extraction from subfigures. (iv) Subfigure clustering.

each subfigure as a *visual word* that captures its low-level image features. Thus, an image based document is described as a sequence of visual words, analogous to a sequence of text terms. The following steps to build an image-based document classifier adapt basic techniques from text document classification. We use a bag-of- words representation and a naïve Bayes classifier.

In previous work [4], we used a combination of supervised and unsupervised learning to define a vocabulary of visual words based on a small subset of the TREC genomics data. In this paper we report an alternative approach, trying to use only unsupervised learning (i.e. pure clustering) to define a vocabulary of visual words based on the whole TREC genomics dataset [2].

2 Document Descriptors via Image Features

Our first task is to represent documents based on their image features. We summarize each subfigure as a visual word, which is essentially the tag of the cluster to which the subfigure is assigned. Table 1 summarizes our representation method.

The details of the steps *figure extraction* (1), *figure segmentation* (2), and *image feature extraction* (3) have been presented in our previous publications [1, 5] and are omitted here. Figure 2(ii) demonstrates the segmentation result for an example figure. For each subfigure, 44 features are computed as illustrated in Figure 2(iii).

In step 4, we use clustering to group together subfigures in the training set that share similar characteristics. There is a large choice of clustering algorithms. Here we use K-Means provided in the Weka toolkit [6] due to its simplicity and efficiency.

Input: The training set* of text documents in XML format.
Output: A vocabulary of visual words and image-based document descriptors.
Steps:
(1) Extract figure URLs from each document, and obtain figure images from the publisher's web site.
(2) Segment each figure into subfigures based on connected components analysis.
(3) Extract low-level image features from each subfigure.
(4) Cluster all the subfigures in the training set.
(5) Assign a cluster tag to each subfigure, and create a vocabulary of visual words.
(6) Create an image-based document descriptor, i.e. a sequence of visual words for each document.
* We note that the class label assigned to each training document is not used in our method of feature extraction, but used in document classification (Section 3).

Table 1. An outline of our method for producing image-based document descriptors.

In step 5, each subfigure is assigned a tag, based on the cluster to which it belongs. This tag is used a surrogate for the figure and is treated as a special word, which we call a *visual word.* The collection of visual words forms the vocabulary for document representation. The number of clusters determines the size of this vocabulary. Ultimately, clustering should group together objects that are more similar to each other than to objects in all other clusters. If the number of clusters is too small, the objects within a cluster may not share much similarity. If the number is too large, objects may be separated into fine clusters despite their similarity.

In step 6, all the subfigures in each document are represented as a sequence of visual words. We choose the five images that are nearest to the cluster centroid as the cluster representatives. For the test set, each subfigure is assigned a cluster tag (a visual word) by finding its nearest neighbor among the representative images of each cluster. The Image-based document description for the document shown in Figure 1 is presented in Table 2.

C774 C881 C962 C431 C998 C957 C517
C476 C476 C384 C333 C899
C659 C899 C990 C438 C778
C892 C868 C853 C921 C892
C370 C416 C993 C766 C695 C766 C741 C737 C836 C147 C316
C300 C902 C182 C786 C963 C759

Table 2. Visual words describing the document shown in Figure 1. The size of the vocabulary is 1000. This document contains 6 figures and 39 subfigures. The second line (shown in bold) contains the visual words corresponding to the figure shown in Figure 4.

3 Image-based Document Classification

We next describe how classification is applied to documents whose representation is based on image features. Document representation using image-based features is adapted from the bag-of-words approach commonly used in text categorization. Each visual word is treated independently. A document d is represented as an n-dimensional vector $d = (d_1, d_2, \ldots, d_n)$, where d_i is the term weight. We use the well-known $tf \cdot idf$ weighting scheme [3], where d_i is proportional to the frequency of the term within the document (*tf*), and inversely proportional to the number of documents containing the term (*idf*).

Once the feature vectors are formed, we build a naïve Bayes Classifier using the Weka toolkit [6]. The naïve Bayes classifier is built by obtaining statistics from the set of labeled training data. A document D is assigned to the class C that maximizes the likelihood: $\Pr(D|C) = \prod_{i=1}^{n} \Pr(d_i|C)$. Expressing the conditional probability $\Pr(D|C)$ as a product of simpler probabilities is based on the (naïve) assumption of conditional independence among the features, given the class.

4 Experiments and Results

We performed our experiments on the biomedical document classification tasks defined by the TREC Genomics track 2005 [2], which included four subtasks, denoted as *G, T, E,* and *A*. Each of them can be viewed as a binary classification task. The documents are categorized as either *relevant* or *irrelevant* for curation.

A total of 5,837 biomedical articles were designated as the training set, while 6,043 articles were used as the test set. The four tasks all use the same set of training and test documents. A document may of course be labeled differently with respect to the different tasks.

For each subtask, we train an image-based classifier and test it on the whole test set. We use the same evaluation metrics used to evaluate submitted runs in TREC Genomics track [2]. The primary evaluation metric is the normalized *Utility* value. Other measures include the standard *precision*, *recall,* and *F-score* (combining recall and precision). Table 3 summarizes our preliminary results when using image-based document classifiers and a vocabulary of 1000 visual words. We also list the median and the minimum results from TREC 05 (which are based on text – not on image data) for an informal comparison. Our current image-based classifier performs below the TREC05 median, but above the minimum with regards to *Utility*.

The focus of our work is not to outperform the text-based systems which participated in TREC 05. A total of 46-48 runs were submitted for each of the four tasks, using a variety of text features. However, none of them use analysis of figure images. The main contribution of this work is the exploration of a new space of features,

based purely on the clustering of subfigures for document classification. This new space of features is aimed to supplement text-based features.

		Precision	Recall	F-measure	Utility
G	Image-based	0.1162	0.4459	0.1844	0.1376
	Trec05 minimum	0.0706	0.1023	0.0979	-0.0342
	Trec05 Median	0.2102	0.6506	0.3185	0.4575
A	Image-based	0.1278	0.5422	0.2069	0.3246
	Trec05 minimum	0.2191	0.25	0.2387	0.2009
	Trec05 median	0.3572	0.8931	0.5065	0.7773
E	Image-based	0.0671	0.6286	0.1212	0.492
	Trec05 minimum	0	0	0	-0.0074
	Trec05 median	0.12195	0.8	0.1985	0.6413
T	Image-based	0.0176	0.55	0.0341	0.4169
	Trec05 minimum	0.0132	0.05	0.026	0.0413
	Trec05 median	0.0526	0.9	0.0952	0.761

Table 3. Classification results, using the evaluation metrics described by Hersh *et al.*[2].

We also tested the effects of modifying the number of clusters used. We notice that the number of clusters affects the classification performance. The effect is not consistent for all the classification tasks. Due to space limitation, we do not present the details here. The evaluation of clustering performance and the choice of an appropriate number of clusters merit further study.

5. Conclusion

In this paper, we propose a method that uses unsupervised clustering to characterize each subfigure in a document as a visual word, and thus create an image-based document description. This description is analogous to that used for text-based representation of documents. We are therefore able to apply the bag-of-words representation and standard classification methods to train an image-based classifier.

We described here the extraction of simple, low-level image features from the subfigures, under the assumption that these image features are useful for document feature representation. A wide variety of features can be extracted from an image. We believe that the choice of features is important for clustering images into meaningful groups. This is a subject we are currently investigating. We are also exploring the incorporation of domain knowledge and supervised learning to describe image features in a semantically meaningful way. Other methods of representing features from the figures in a document are being studied as well.

It is important to note that the image-based classifier is not meant to replace, but to assist in text-based classifier. Based on our preliminary experiments in this area [4], we expect that combining image and text analysis will help resolve ambiguity and improve the effectiveness of literature mining.

Acknowledgements

We gratefully acknowledge support from Canada Natural Sciences and Engineering Research council (grants 298292 and 41635) and the Xerox Foundation, and by the CFI New Opportunities award 10437.

References

[1] N. Chen, H. Shatkay, and D. Blostein. Use of Figures in Literature Mining for Biomedical Digital Libraries. *Proc. of the 2nd IEEE Int. Conf. on Document Image Analysis for Libraries (DIAL'06).* pp. 180-197.

[2] W.R. Hersh, A. Cohen, J. Yang, R.T. Bhuptiraju, P. Roberts, M. Hearst. TREC 2005 Genomics Track overview. *Proc. of TREC 2005*, NIST Special Publication, 2006. pp. 14-25.

[3] F. Sebastiani. Machine Learning in Automated Text Categorization. *ACM Computing Surveys* 34(1). 2002. pp. 1-47.

[4] H. Shatkay, N. Chen, D. Blostein. Integrating Image Data into Biomedical Text Categorization. Proc. Of the *Int. Con. on Intelligent Systems for Molecular Biology* (ISMB) *2006*. pp. e446-e453.

[5] H.R. Widlund, M.A. Horstmann, E.R. Price, et al. Beta-catenin-induced melanoma growth requires the downstream target Microphthalmia-associated transcription factor. *J. of Cell Biology.* 158(6). 2002. pp. 1079-87.

[6] H. Witten and E. Frank. *Data Mining: Practical machine learning tools and techniques.* Morgan Kaufmann. 2005.