Research Paper Classification Using Link-Based Classification and Citation Contexts

DREU Final Report

Namchi Do

Abstract

Text classification is a machine-learning task where the goal is to label a document based on its attributes. Automating text classification is an important task because there is a plethora of unclassified data, and classifying text manually is both expensive and time-consuming. This project employs a combination of two previously proposed methods of classification. The first is the use of a document's links, which represent how the document is related to other documents in the network. Exploiting this relationship between documents can tell us more about a particular document since elements of a network influence each other. Secondly, this project incorporates citation contexts. Although citation contexts are not commonly found in research papers, they serve as micro-summaries of a paper and can be used to more accurately classify a document. Thus, by combining these two approaches to classification, we hope to see more accurate classification results.

Introduction

Text classification is the task of automatically assigning a label to a document based on textual information and statistics that can be derived from the text.

Text classification is an important task because it promotes efficiency. For instance, users save time and are more likely to find what they are searching for if documents are properly classified. Classified data is easier to manage and retrieve. However, the process of manually labeling data is inefficient, time-consuming, and error-prone. Thus, text classification is a way to automate the process.

There are many real-world applications of text classification, such as spam email filtering, document genre categorization, and image content classification. Researchers have also focused on web page classification augmented by hypertext data (Blum & Mitchell, 1998), research paper classification, and keyphrase extraction. Solutions to classification problems can also be extended for a variety of purposes, such as collaborator suggestions in a network, automatic keyphrase tagging, or recommended reading suggestions.

The question that this project addresses is whether we can come up with a relatively accurate classification algorithm by combining two recent studies that produced better-than-the-standard accuracy for classification. The first makes use of link diversity to exploit the relational structure of a research paper network (Lu & Getoor, 2003) and the second uses citation contexts (Gollapalli & Caragea, 2014). The data set used is a collection of research papers from CiteSeerX.

The following sections will discuss relevant literature, terminology, the algorithm, the results of experiments, and end with the discussion.

Related Works

Text classification techniques have been applied to several different forms of text, including web pages, email, research papers, and tweets. In general, the steps for text classification are: (1) predetermining the classes to which a document can belong, (2) learning the rules for each class, (3) training classifiers based on those rules, and, finally, (4) predicting the classes of the documents in the test set. Classification algorithms fall into one of three categories: supervised, unsupervised, and semi-supervised.

Supervised methods of text classification build classifiers directly based on pre-labeled data. Supervised algorithms are the most common in text classification, and many solutions have been proposed, including the Naive Bayesian approach (Rocchio, 1971), K-Nearest Neighbor (Yang, 1999), Support Vector Machines (Joachims, 1997), and recently, link-based classification (Lu & Getoor, 2003), which this project is based upon. Although supervised methods generally perform better than

unsupervised and semi-supervised solutions, the caveat is that they require a sizeable quantity of labeled data, which can be difficult to obtain.

One advantage unsupervised algorithms have over supervised ones is that it is much easier to collect unlabeled data; this circumvents this need for labeled training data. Previous approaches to unsupervised methods include Expectation Maximization (Blum & Mitchell, 1998) and text clustering (Martinez-Romo, Araugo, & Duque, 2015). The idea behind unsupervised approaches is to extract the hidden structure in the unlabeled data through statistical approaches.

Semi-supervised, or partially supervised, text classification combines elements of the both supervised and unsupervised methods. For instance, a semi-supervised approach might make use of a small set of labeled data augmented by a large set of unlabeled data. This was the premise behind Blum and Mitchell's co-training approach, which had "significant improvement" in practice for web page classification (Blum & Mitchell, 1998). Likewise, in their experiments with document classification, Liu et al. showed that having some labeled data for positive examples and only unlabeled data for negative examples was enough to build accurate classifiers (Liu, Lee, Yu, and Li).

A closely related problem that also formed the basis of this research project is the task of keyphrase extraction. The goal of keyphrase extraction is to generate a set of potential keywords or keyphrases to summarize the main topics of a text. Like text classification, the algorithms for keyphrase extraction can be supervised, unsupervised, or semi-supervised.

In supervised keyphrase extraction, documents are tagged with user-input or author-input keyphrases, which are then used to train the classifiers. Unsupervised methods rely on statistical feature such as phrase frequency, position of first occurrence, and term frequency-inverse document frequency to determine how likely a candidate keyphrase is to be a keyphrase (Caragea, Bulgarov, Godea, & Gollapalli, 2014). Several unsupervised methods are graph-based, including PageRank, TextRank, CiteTextRank (Gollapalli & Caragea, 2014), and SemGraph (Martinez-Romo, Araugo, & Duque, 2015).

This research project relies mainly on two prior experiments: the iterative classification algorithm (ICA) proposed by Lu & Getoor (2003), and Caragea & Gollapalli's incorporation of citation contexts for predicting keyphrases for research papers (2014).

The ICA proposed by Lu and Getoor is a supervised approach to text classification. One of the novel features of the algorithm is the incorporation of link distributions to model the dependencies in a network. Lu and Getoor define links as follows (2003):

> L – the set of links between objects in a network
>
> $I(X_i)$ – the set of incoming edges for X_i such that $\{ X_j \mid L_{j \to i} \in L \}$
>
> $O(X_i)$ – the set of outgoing edges of X_i such that $\{ X_j \mid L_{i \to j} \in L \}$
>
> $Co(X_i)$ – the set of objects co-cited with X_i such that $\{ X_j \mid X_i \neq X_j$ and $\exists X_k$ that links X_i and $X_j \}$

In other words, an element in the set $I(X_i)$ is a document that X_i references or cites. An element in $O(X_i)$ is a document that references or cites X_i. Documents in the set $Co(X_i)$ include ones that reference a same source as X_i, are referenced by a source that also references X_i, etc.

The significance of the ICA is that it exploits the structure of a network. The assumption is that objects that are linked in a network influence each other, so modeling the class distributions of neighboring objects can help predict a target document's class. The conclusion of this ICA experiment was that using link distributions improved classification accuracy (Lu & Getoor, 2003).

The second line of research that motivates this research project is the use of citation contexts as a classification feature. Citation contexts

refer to the short description of the cited source and are essentially summaries of the source as well as the paper that contains it. Experiments that incorporated citation networks, both in supervised (Caragea, Bulgarov, Godea, & Gollapalli, 2014) and unsupervised (Gollapalli & Caragea, 2014) approaches to keyphrase extraction, showed that the additional information from citation contexts improved accuracy.

Thus, this research project will combine Lu and Getoor's ICA algorithm and the use of citation network contexts to see if this combination can improve research paper classification.

Terminology

This project looks at research papers and defines terms as follows:

Cited(X_i) – the set of papers which reference paper X_i. Thus, a paper X_j falls into this set if X_i is cited by X_j.

Citing(X_i) – the set containing papers that X_i references. A paper X_j falls into this set if X_i cites X_j.

Co-Cited(X_i) – the same set as Lu and Getoor's definition of *Co*(X_i)

Papers are classified based on different features, which includes several different contexts and the link diversity:

Global context – the paper's title and abstract

Cited context – the citation contexts, or descriptions, for papers that cite the target paper

Citing context – the citation contexts for papers that the target paper cites

Citation context – both the cited and citing contexts combined

Link diversity – the frequency counts of the class labels of the documents in *Cited*(X_i), *Citing*(X_i), and *Co-Cited*(X_i) for a target document X_i

Algorithm & Experiment

The algorithm proposed for this project is as follows.

[1] *Preprocessing* – this includes:
- Generating a citation network dictionary formatted as [p_1, p_2] where p_1 is a paper cited by p_2
- Splitting the research papers into several sets for training and testing
- Generating input WEKA Instances to model the global context, citation (or cited/citing) context, and link diversity of the training sets and only the global context and citation (cited/citing) context for the test sets.

[2] *Training* – train the classifiers on the input training Instances. Separate classifiers are trained on the global context, citation (cited/citing) context, and link diversity.

[3] *Bootstrap* – this comprises:
- Assigning an initial class to each document in the test set using the global context and citation context classifiers
- Generating a link diversity Instance for the test set based on the current classifications

[4] *Iterative classification*
- Recompute the link diversity statistics based on current classifications
- Reclassify the document based on the results from the global context, citation context, and link diversity classifiers
- Repeat this step until reaching a stopping point. For this project, the stopping conditions are when (1) there no more changes to the classification, or (2) the reclassification process enters a loop, such as when the

same documents alternate their class predictions

Steps [3] and [4] are modified from Lu and Getoor's ICA, where the stages were termed "Boostrap" and "Iteration," respectively.

The original dataset for this project is a bank of 3,186 research papers from CiteSeerX. After removing papers that did not have both a citing and cited context, there were 2,431 papers left. From these papers, ten cross-validated sets were produced.

Three WEKA classifiers were used for each set: Naïve Bayes, SMO, and J48 Decision Tree. Due to timing constraints, the Simple Logistic Regression classifier was not tested. Data for the accuracy, precision, recall, and f-measure was collected at several stages. The equations for these are:

Accuracy = $(TP + TN) / (P + N)$

Precision = $TP / (TP + FP)$
Recall = $TP / (TP + FN)$
F-measure = $2 * (Precision * Recall) / (Precision + Recall)$

Where P is the number of positives, N negatives, TP true positives, FP false positives, TN true negatives, and FN false negatives.

Experiments were conducted with the full citation context, then only the cited context, and then only the citing context in the model, to see which would produce the best results. In each, the final classifications depended on a weighted average of the global context, citation (or cited/citing) context, and current classifications given to the link diversity.

Results

The statistics after the Bootstrap stage and final stage are:

	Global Context	Citation Context	Cited Context	Citing Context	Link Diversity
Avg. Accuracy	0.741	0.767	0.712	0.766	0.69
Avg. Precision	0.738	0.771	0.72	0.766	0.7
Avg. Recall	0.741	0.767	0.712	0.766	0.69
Avg. F-measure	0.736	0.766	0.712	0.762	0.686

Table 1. Statistics after Bootstrap for Naïve Bayes Multinomial Classifier.

	Global Context	Citation Context	Cited Context	Citing Context	Link Diversity
Avg. Accuracy	0.593	0.607	0.542	0.584	0.641
Avg. Precision	0.588	0.602	0.534	0.58	0.638
Avg. Recall	0.593	0.607	0.542	0.584	0.641
Avg. F-measure	0.586	0.602	0.534	0.578	0.634

Table 2. Statistics after Bootstrap for J48 Classifier.

	Global Context	Citation Context	Cited Context	Citing Context	Link Diversity
Avg. Accuracy	0.661	0.681	0.51	0.656	0.618
Avg. Precision	0.66	0.681	0.526	0.655	0.676
Avg. Recall	0.661	0.681	0.51	0.656	0.618
Avg. F-measure	0.658	0.675	0.504	0.652	0.609

Table 3. Statistics after Bootstrap for SMO Classifier.

	Citation Context	Cited Context	Citing Context
Avg. Accuracy	0.704	0.705	0.706
Avg. Precision	0.712	0.713	0.713
Avg. Recall	0.704	0.705	0.706
Avg. F-measure	0.701	0.702	0.703

Table 4. Statistics after last stage (Iterative Classification) for Naïve Bayes Classifier.

	Citation Context	Cited Context	Citing Context
Avg. Accuracy	0.675	0.681	0.676
Avg. Precision	0.671	0.679	0.672
Avg. Recall	0.675	0.681	0.676
Avg. F-measure	0.667	0.674	0.668

Table 5. Statistics after last stage (Iterative Classification) for J48 Classifier.

	Citation Context	Cited Context	Citing Context
Avg. Accuracy	0.641	0.638	0.641
Avg. Precision	0.688	0.685	0.687
Avg. Recall	0.641	0.638	0.641
Avg. F-measure	0.631	0.628	0.631

Table 6. Statistics after last stage (Iterative Classification) for SMO Classifier.

Discussion

Initially in the Bootstrap stage, all three classifiers showed that the citation context is more accurate than the citing and cited contexts, and that the citing context is more accurate than the cited. This implies that a research paper is more accurately described in its citing paper. All three papers also show that the full citation context is more accurate than the global context. This part of the result validates the earlier finding by Caragea, Bulgarov, Godea, and Gollapalli that citation contexts can improve classification accuracy (2014).

Of the three classifiers, the Naïve Bayes was overall the most accurate. Its final f-measures were over 3% more accurate than the final J48 f-measures. All three classifiers produced results that were more accurate than the original link diversity results.

However, there were some unexpected results that seem to conflict earlier findings, leading to the possibility that this experiment had a few bugs.

For the Naïve Bayes and SMO, the final statistics were worse than the initial statistics for the global and citation contexts, which implies that the link diversity here did not improve classification. These results are the opposite of what Lu and Getoor found using their logistic regression model (2003). Thus, the results from this experiment are not conclusive enough to draw a strong conclusion.

Future Work

There are still many ways to improve this project. More experiments should be done and the algorithm must be more closely examined for bugs that would have skewed the results. The algorithm in this project should also try a logistic regression model to see how it compares with Lu and Getoor's results.

Acknowledgements

This project was funded by the Computing Research Association-Women Distributed Research Experiences for Undergraduates (CRA-W DREU) program. Special thanks to Dr. Cornelia Caragea from the University of North Texas for providing the datasets and lots of guidance.

References

Blum, A., & Mitchell, T. (1998). Combining labeled and unlabeled data with co-training. *COLT: Proceedings of the eleventh annual conference on Computational learning theory.*

Caragea, C., Bulgarov, F., Godea, A., & Gollapalli, S. D. (2014). Citation-enhanced keyphrase extraction from research papers: A supervised approach. EMNLP.

Gollapalli, S. D., & Caragea, C. (2014). Extracting keyphrases from research papers using citation networks. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence* (pp. 1629-1635).

Joachims, T. (1998). Text categorization with Support Vector Machines: Learning with many relevant features. *Machine Learning: ECML-98, Tenth European Conference on Machine Learning*.

Lu, Q., & Getoor, L. (2003). Link-based classification. *ICML 2003 Proceedings of the Twentieth International Conference on Machine Learning*.

Martinez-Romo, J., Araujo, L. & Duque Fernandez, A. (2015), SemGraph: Extracting keyphrases following a novel semantic graph-

based approach. *Journal of the Association for Information Science and Technology.*

Rocchio, J. (1971). Relevance feedback in information retrieval. In G. Salton (Ed.), *The SMART retrieval system: experiments in automatic document processing. Englewood Cliffs, NJ: Prentice Hall.*

Yang, Y. (1999). An evaluation of statistical approaches to text categorization. *Journal of Information Retrieval.*