The Use of Active Learning in Text Categorization

Ray Liere
lierer@mail.cs.orst.edu

Prasad Tadepalli
tadepall@research.cs.orst.edu

Department of Computer Science, Oregon State University,
Dearborn Hall 303, Corvallis, OR 97331-3202, USA

Abstract

With the advent of large distributed and dynamic document collections (such as are on the World Wide Web), it is becoming increasingly important to automate the task of text categorization. The use of machine learning in text categorization is difficult due to characteristics of the domain, including a very large number of input features, noise, and the problems associated with semantic analysis of text. As a result, the use of supervised learning requires a relatively large number of labeled examples. We explore the possibility of using (almost) unsupervised learning and propose some novel approaches to using machine learning in this domain.

1. Introduction

The area of information access can be usefully divided into text categorization and information retrieval. In the text categorization task, there are a finite number of categories which are known in advance. The learning program's task is to learn a set of rules that classify documents into their appropriate categories. The task in information retrieval is to respond to queries with documents that satisfy the query. Learning in information retrieval is much more challenging than in text categorization because of the richness of the query language. In this paper, we consider a variety of learning problems that occur in text categorization, the area of our initial experimental studies.

Text categorization is difficult due to certain characteristics of the domain: text is not naturally represented as a feature vector; even when it is, there is a large number of features; there is a large number of documents; there is potentially a large variation in the amount of information in each document; while we are interested in the concepts the documents contain, the documents are written using words; the documents are written in a natural language and so may contain ambiguities; the information needed to correctly classify a document may be completely implicit or hidden (depending on the representation used); the documents are written by humans and so may contain errors; the documents may contain a host of tokens besides words, such as numeric information, abbreviations, etc. High dimensionality and noisy data are the main domain characteristics that make machine learning in this domain very challenging.

This research was partially supported by the National Science Foundation under grant number IRI-9520243.

Most existing methods of text categorization fall into one of 3 categories: boolean, probabilistic, or vector space [Fung95]. Existing methods use either a knowledge engineering or a machine learning technology. Admittedly there are many situations when the dividing line between methods or technologies is fuzzy at best, and many approaches draw from multiple methods and technologies. Most existing methods that use machine learning, even in the broadest sense of the term, use some form of supervised learning.

2. Text Categorization

Much progress has been made in the past 10-15 years in the area of text categorization and in applying machine learning to text categorization. We seem, however, to have reached a plateau (albeit a reasonably high one). It is often difficult to detect statistically significant differences in overall performance among several of the better systems, whether one is employing knowledge engineering or supervised machine learning. One often finds comparisons being made on the basis of fractions of a percentage point difference in some performance metric. Many methods, although quite different in the technologies used, seem to perform about equally well overall.

One of the most impressive results in applying machine learning to text categorization is by Apté and Damerau, who used optimized rule-based induction and reported an 80.5% recall-precision breakeven point using the Reuters-22173 collection [Apte94]. However, to achieve this result, over 10,000 labeled examples were used.

Typically, large numbers of labeled examples, usually in the thousands, are needed in order to obtain good results with supervised learning. Knowledge engineering methods, on the other hand, require large amounts of engineering and usually large amounts of computational effort on the front end.

2.1 Labeled versus Unlabeled Examples

Castelli and Cover use Bayesian analysis to determine the relative worth of labeled versus unlabeled examples [Castelli95]. They conclude that labeled examples are exponentially more valuable than unlabeled examples. This indicates that one will save a great deal of (computational) learning effort by using labeled rather than unlabeled examples.

Unfortunately, examples do not label themselves – a human must do that. For some situations, this is not practical and perhaps not even possible. The "exponentially more valuable" aspect of labeled examples can be viewed as the result of partially solving the learning problem by using a preprocessor that utilizes resources, such as time and money, and in this case happens to be a human being. There is thus a tradeoff between the human doing the labeling so that the computer can do less work learning (supervised learning) and the human doing "no" work and the computer spending more time learning (unsupervised learning).

2.2 Unsupervised Learning

A natural conclusion to draw is that one ought to use unsupervised learning. This would allow the use of unlabeled examples, which are certainly more plentiful than labeled ones. The dream is to feed all of the full text documents in the collection into an unsupervised learning algorithm, and it will perform text categorization. Of course, the question is, will the categorization determined by the algorithm be a categorization that will be of some use to a human? There are many reasonable ways to categorize a particular collection of documents. There is no reason to think or even hope that categorization by general topic content is how an unsupervised learning algorithm will do it, in the absence of any additional information, suggestions, proddings, etc.

We have started conducting some experiments with AutoClass C, an unsupervised Bayesian classifier [Cheeseman95], using newspaper articles from the Reuters-22173 corpus. While our experiments have just started, we have already found that blind unsupervised learning, when applied to documents viewed as bags of words, takes a lot of real and virtual computer memory, a lot of time, and will not necessarily find the categories that the human observer thinks are important. This is emphatically *not* a criticism of AutoClass C. We are using AutoClass C in a domain in which it is not normally used. Most problems on which AutoClass C is used deal with far fewer features [Cheeseman95]. Also, when using AutoClass C on a large and complex data set, quite a bit of effort should be expected to be spent in cycling through the tasks of data analysis, data exploration, data transformation, and the inspection of results.

2.3 The Bayesian Network Formalism

We selected AutoClass C for many reasons. One of the most important was that it can be seen as an unsupervised learning algorithm for a particular instantiation of the Bayesian network formalism, namely a 2-layered Bayesian network. We expected from early on to be faced with the situation we are in now – that of having to modify the system so that it will better perform the task of text categorization. We wanted a firm basis on which to build.

The Bayesian network formalism has a strong theoretical (as well as empirical) foundation [Pearl88]. It is quite general and is much more efficient to use than the exhaustive examination of the full probability distribution in situations where some degree of locality applies.

Work using Bayesian networks in text categorization has been done by (1) Fung, Crawford, Appelbaum, and Tong [Fung90]; (2) Fung and Del Favero [Fung95]; (3) Turtle and Croft [Turtle90]; and (4) Haines and Croft [Haines93]. These approaches typically use a Bayesian network with an at least partially predefined structure, and then employ supervised learning to determine the structure details and the probability distributions of the nodes in the network. However, we are attempting to use an unsupervised Bayesian classifier to learn the structure of the Bayesian network as well as the probability distributions.

2.4 A Little Bit of Supervision

We feel that an area warranting further research is the use of "some" supervision. We use the term "active learning" to refer to any form of learning whereby the learning algorithm has some degree of control over the inputs on which it is trained.

If one were to limit user input, then one could perhaps compromise between the disadvantages of supervised and unsupervised learning. Many examples labeled by the expert in the supervised setting are in fact redundant in the sense that any reasonable hypothesis can be learned from a much smaller number of labeled examples, if only they are carefully selected.

Recently there have been some promising results in the active learning area by: (1) Board and Pitt (semi-supervised learning) [Board87]; (2) Freund, Seung, Shamir, and Tishby (user selection of examples for labeling) [Freund92]; (3) Lewis and Gale (uncertainty sampling) [Lewis94]; (4) Cohn, Atlas, and Ladner (selective sampling) [Cohn94]; and (5) Dagan and Engelson (committee-based sampling) [Dagan95]. While approaches and results vary, all of these studies concluded that these various forms of active learning improved learning efficiency by significant amounts.

Results to date seem to suggest that the addition of active learning capabilities to a generalized Bayesian classifier such as AutoClass C may produce improved results when applied to text categorization. By decreasing the effort required in the labeling of examples, large distributed dynamic document collections such as those on the World Wide Web could be cost-effectively categorized by the computer.

3. Other Areas to Explore

There are at least 2 other areas that one can explore. In fact, these are fairly independent of the incorporation of active learning, and hence one or both of them may be combined with it.

3.1 Learn Multi-Layered Bayesian Networks

AutoClass C assumes that there are only two layers in the network, an unobserved layer (the categories) and an observed layer (the documents – in most cases, words). A two-layer Bayesian network algorithm that computes the best solution would probably not be practical with this domain, especially with the large number of features and the large number of documents. It would seem that the accuracy of text categorization could be improved if the number of layers in the Bayesian network learned by AutoClass C could be increased to at least 3, with the intermediate layers being used to represent phrases or even concepts.

3.2 Accommodate Multiple-Category Documents

Most documents belong to more than one category. However, most unsupervised conceptual clustering algorithms partition the data instances into classes. We are only familiar with two that do not have this property – OLOC [Martin94] and UNIMEM [Lebowitz87]. Supervised learning gets around this problem by effectively training a system for each category. For each category, the system gives a "yes" or "no". The fact that a document can be in more than (or less than) one category does not ever need to be directly confronted.

4. Conclusions

The use of machine learning in text categorization is both challenging and promising. Existing supervised and unsupervised learning methods seem to suffer from a number of serious problems when applied to this task. We advocate an approach that is intermediate between supervised and unsupervised learning and that involves active learning. We believe that the Bayesian network formalism provides a good framework for this task. We plan on researching and implementing the addition of active learning capabilities to a generalized Bayesian learner such as AutoClass C. We are also planning to extend AutoClass C to make it applicable to more general kinds of tasks such as overlapping categories and multi-layered Bayesian networks. These are all challenging problems.

5. Acknowledgements

The availability of AutoClass C [Cook] and the Reuters-22173 corpus [Reuters] has greatly assisted in our research to date.

6. References

[Apte94] Chidanand Apté, Fred Damerau, Automated Learning of Decision Rules for Text Categorization, *ACM TOIS* 12(2):233-251, July 1994

[Board87] Raymond A. Board, Leonard Pitt, Semi-Supervised Learning, Department of Computer Science, University of Illinois at Urbana-Champaign, Report No. UIUCDCS-R-87-1372, September 1987

[Castelli95] Vittorio Castelli, Thomas M. Cover, On the Exponential Value of Labeled Samples, *Pattern Recognition Letters* 16(1):105-111, January 1995

[Cheeseman95] Peter Cheeseman, John Stutz, Bayesian Classification (AutoClass): Theory and Results, in *Advances in Knowledge Discovery and Data Mining,* The AAAI Press: Menlo Park, expected March 1996

[Cohn94] David Cohn, Les Atlas, Richard Ladner, Improving Generalization with Active Learning, *Machine Learning* 15(2):201-221, May 1994

[Cook] AutoClass C, version 2.7, software and documentation; Diane Cook, Joseph Potts, Will Taylor, Peter Cheeseman, and John Stutz; available via ftp from: `csr.uta.edu:/pub/autoclass-c.tar.Z`

[Dagan95] Ido Dagan, Sean P. Engelson, Committee-Based Sampling for Training Probabilistic Classifiers, ML95, 1995, p. 150-157

[Freund92] Yoav Freund, H. Sebastian Seung, Eli Shamir, Naftali Tishby, Information, Prediction, and Query by Committee, NIPS92, p. 483-490

[Fung90] Robert M. Fung, Stuart L. Crawford, Lee A. Appelbaum, Richard M. Tong, An Architecture for Probabilistic Concept-Based Information Retrieval, SIGIR'90, 1990, p. 455-467

[Fung95] Robert Fung, Brendan Del Favero, Applying Bayesian Networks to Information Retrieval, *Communications of the ACM,* 38(3):42-48, March 1995

[Haines93] David Haines, W. Bruce Croft, Relevance Feedback and Inference Networks, SIGIR'93, p. 2-11

[Lebowitz87] Michael Lebowitz, Experiments with Incremental Concept Formation: UNIMEM, *Machine Learning* 2(2):103-138, 1987

[Lewis94] David D. Lewis, William A. Gale, A Sequential Algorithm for Training Text Classifiers, SIGIR'94, p. 3-12

[Martin94] Joel D. Martin, Dorrit O. Billman, Acquiring and Combining Overlapping Concepts, *Machine Learning* 16(1-2):121-155, July/August 1994

[Pearl88] Judea Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference,* Morgan Kaufmann, 1988

[Reuters] Reuters-22173 corpus, a collection of 22,173 indexed documents appearing on the Reuters newswire in 1987; Reuters Ltd, Carnegie Group, David Lewis, Information Retrieval Laboratory at the University of Massachusetts; available via ftp from: `ciir-ftp.cs.umass.edu:/pub/reuters1/corpus.tar.Z`

[Turtle90] Howard Turtle, W. Bruce Croft, Inference Networks for Document Retrieval, SIGIR'90, p. 1-24