Classifying Text Documents using Modular Categories and Linguistically Motivated Indicators.

Eleazar Eskin
eeskin@cs.columbia.edu
Department of Computer Science
Columbia University
New York, NY 10027

Matt Bogosian
mattb@cs.columbia.edu
Department of Computer Science
Columbia University
New York, NY 10027

Abstract

In this paper we present two improvements to traditional machine learning text classifiers. The first improvement we present is a decomposition of the classification space into several dimensions of categories. This breaks down the categorization problem into smaller more manageable parts. We discuss when decomposition is useful. The second improvement is to incorporate linguistically motivated indicators to supplement the classification. These indicators provide information about the structure of the document which are used to improve the classification accuracy.

Introduction

The World Wide Web is perhaps the most challenging environment for text classification. Because of its heterogeneity, the Web contains many documents which are inherently difficult to categorize. Applications which can make use of text classification include summarization systems and Web crawlers.

Machine Learning text classifiers on the other hand have become effective in classifying documents based on keywords into a small set of categories. However, the performance of these classifiers decreases with the number of categories into which they classify. In addition, because classifiers typically rely on only word frequencies in a document, the classifiers are susceptible to misclassifying documents where the classification problem is primarily dependent on structural features of the document. For example, in distinguishing between interviews and press reports on the same issue, the format of the interview helps categorize the document. These inadequacies cause the classifiers to become less useful when applied to the World Wide Web.

Previous Work

Previous relevant work in genre identification and text classification have come from two communities including the machine learning community, and the natural language processing community.

In the machine learning community, Tom Mitchell's Naive Bayes classifier uses only the words of the document as the features for the machine learning algorithm.

In the computational linguistics community, linguistically based numerical features are used for genre identification (Biber 89). Most of the work revolve around statistical analysis of document features and identifying their correlations to genre. Biber also performed analysis over multiple features for a document, but not in order to decompose the category space. Kessler applied machine learning algorithms to these features to create classifiers.

Larkey and Croft (1996) combine classifiers for categorization, but their classifiers are different algorithms applied to the same category scheme, while this work focuses on applying the same classification algorithm to different dimensions or "parts" of the categorization scheme.

Improvements to Traditional Machine Learning Text Classifiers

Two improvements to machine learning text classifiers are discussed here. The first improvement is a decomposition of the category space into dimensions in order to increase the number of categories that a classifier can handle. This type of approach is common in the machine learning community. A typical example of an analogous approach applied to a different problem would be using multiple classifiers over images. (Mitchell 97)

The second improvement is incorporating linguistically motivated features into the classification. This approach builds upon work in the Natural Language Processing community on genre identification. (Biber 89)

Dimensions of Classification

The classification scheme of documents can be decomposed into several independent parts in order to increase the overall accuracy of the classification. We do this by partitioning the classification category space into orthogonal dimensions and building a separate classifier for each dimension.

This type of approach can be applied to many classifications. For example, in his text on *A Typology of English Texts*, Biber uses 23 categories for genre, 17 of which are genre for written text.

Press reportage
Editorials
Press Reviews
Religion
Skills and Hobbies
Popular Lore
Biographies
Official Documents
Academic Prose
General Fiction
Mystery Fiction
Science Fiction
Adventure Fiction
Romantic Fiction
Humor
Personal Letters
Professional Letters

A machine learning classifier would have to learn how to distinguish between all 17 categories. However, we notice that the problem can be partitioned into several parts. We can first attempt to classify the documents by the "type" of the document, i.e. Fiction, Letter, Press Reviews, Editorials, etc. Likewise, we can attempt to classify the "topic" of the document, i.e. Academic, General, Mystery, Science, Official/Professional, etc. What remains is a decomposition of the first classification problem into two simpler subproblems with fewer categories to distinguish.

More formally, let us assume that we can decompose a classification category space into two separate dimensions of size N and M. The original classification category space contained $N \times M$ categories. This means that a classifier on the original category space would have to be able to distinguish between $N \times M$ categories, while a classifier over the new space would contain two classifiers that would have to be able to distinguish between N and M categories respectively. Since classifiers perform significantly better when classifying over a smaller set of categories (mostly because there is more room for error) assuming that a decomposition exists the decomposed classifier should in theory perform better than the original flat classifier. Even though both classifications have to be correct in the decomposition scheme, the decomposed classifiers usually perform better than the original classifier because of the discrepancy in performance depending on number of categories.

There are two assumptions which are implicit in the decomposition. They both are related to the orthogonality of the category space. The first assumption is that for every state in the decomposition there is a unique state in the original category space, and for every state in the original category space, there is a unique state in the decomposition space. The second assumption is that the classification tasks of the dimensions are independent. The classification of a single dimension will not depend on other dimensions. Given these two assumptions, we can infer that the classification in the dimension space will have higher accuracy than the classification in the original space. The accuracy is only higher when the cumulative error between the multiple classifiers is lower than the error of a single classifier.

An example where this decomposition is possible is the newspaper article domain. The two dimensions that we can decompose the space into are "article type" and "article topic". The "type" dimension can contain classifications like: "breaking news", "feature article", "editorial", "opinion", etc. The "topic" dimension can contain classifications like: "international news", "business news", "sports news", "political news", etc. This decomposition is completely orthogonal, i.e. every possible combination of the two dimensions are valid, thus the first assumption is satisfied. We also are assuming that distinguishing between a certain "type" does not depend on the "topic" of the document and distinguishing between a certain "topic" does not depend on the "type" of the document. Classifiers over these dimensions have to distinguish between a much smaller number of categories than a classifier without dimensions. Thus the task of classifying news articles can be made easier by using decomposition of the category space.

This technique can be easily extended to more than two dimensions.

A weaker assumption is that every state in the dimension space maps to a state in the category space, but more that one state in the dimension space can map to a state in the category space. For example, if we extend our example above to incorporate the type "movie review", and our original category space contained the category "movie review", there will be states in out dimension space such as ("movie review", "international") and ("movie review","politics"), etc. Each of these states would map to "movie review" in the original category space. This makes the dimension space larger than the original category space. However, this allows some degree of non orthogonality in the decomposition that is more suitable to real tasks.

Linguistically Motivated Features

Linguistically motivated features can be incorporated into a classifier in order to represent information on the "structure" of the document.

The linguistically motivated features can provide information about the document structure to help distinguish between categories when the words within the documents are similar to multiple categories.

The linguistically motivated features include: document length, average sentence length, pronoun usage and punctuation usage. The intuition behind these features is that different categories have different ranges associated for these features. (Karlgren 94)

The feature values that correspond to a category are learned by a machine learning algorithm that takes as input the category classifications and the linguistic features computed over documents in that category.

Combining Classifiers and Indicators

Both of these improvements involve combining output of either classifiers or linguistic indicators to obtain a final output. There are several approaches to combining this information.

For dimensions, we could construct a mapping from the dimension category space to the original category space. We can also use machine learning techniques to create this function implicitly from the training data. The output of the classifiers can be the input to a machine learning algorithm that is trained using the document's classifications in the original category space. Although the machine learning algorithm introduces some variance in this mapping, the advantage to using it is that the system can be trained simply by classifying the documents and without anything else being done manually. For the purposes of the evaluation of this technique, for simplicity, we construct the mapping manually.

For the linguistic indicators, we have to combine that information with the output of the classifiers. One way to do this would be to use the output of classifiers and the linguistic indicators and train using another machine learning algorithm. This way, the linguistic data is combined into the classification automatically.

These techniques work well together because the problem of decomposition requires recombination of the separate classifiers.

Methodology

In order to evaluate these improvements to the classification scheme, systems that used the above mentioned improvements were compared against a control experiment on a unmodified text classifier to obtain a baseline.

A corpus of 300 of manually classifier documents was created and used throughout all of the experiments in this section. These documents were grouped into 11 heterogeneous categories. The categories were chosen to reflect a wide range of types in order to test performance. Some types are very distinct (such as personal home pages and sports news) while other types are much more closely related such as international news and business news.

The categories are:

Business Front Pages
Personal Home Pages
Sports News
Editorials
Movie Reviews
Arts Features
Book Reviews
International News
Business News
Science News
Political News
Other

The classifier goal was to classify the document into the correct category.

Classifiers Tested

Each of the classifiers have the same input and output. They each take as input an unclassified document, and output the category associated with the document. The word based classifier that was used for these experiments is the Naive Bayes classifier. The specific implementation is the RAINBOW implementation from Tom Mitchell's group. When the category space was decomposed into dimensions, multiple RAINBOW classifiers were used, one on each dimension.

The baseline classifier used for evaluation was a flat Naive Bayes classifier which used only the word features. The multiple category dimension classifier used several text classifiers that were trained to separate articles along a certain dimension. The classifiers results were combined using another machine learning algorithm. The text classifiers used were a set of Naive Bayes classifiers. The results were combined using a manually constructed mapping. This mapping takes output of the Naive Bayes classifiers and relates them to the appropriate original category.

The categories were decomposed into the following dimensions:

1. Type
   (a) News
   (b) Business Web Pages
   (c) Personal Home Pages
   (d) Editorials
   (e) Movie Reviews
   (f) Book Reviews
   (g) Other
2. Topic
   (a) Sports
   (b) Arts
   (c) International
   (d) Business
   (e) Science
   (f) Politics

The classifier incorporating linguistically motivated indicators used the output of a Naive Bayes classifier and the computed linguistic features of the document as inputs to a decision tree which outputs the final classification. A decision tree was used in order to preserve the ordering information of the numerical indicators.

The linguistic indicators that were computed were:

Document Length
Average Sentence Length
Ratio of Pronouns per Word
Ratio of First Person Pronouns per Pronoun
Ratio of Third Person Pronouns per Pronoun
Ratio of High Level References Per Pronoun (this, that, etc.)
Ratio of Punctuation per Word
Ratio of Periods per Punctuation
Ratio of Exclamation Marks per Punctuation

Ratio of Question Marks per Punctuation
Ratio of Quotation Marks per Punctuation
Ratio of Colons per Punctuation
Ratio of Semicolons per Punctuation
Ratio of Commas per Punctuation

These indicators are very similar to the ones in (Kessler 97).

Results

The number of categories is 11. Randomly guessing categories would give 9% accuracy. The flat Naive Bayes classifier (the baseline) gives 68.24% accuracy. The classifier for dimensions gives the following accuracies. For types, the accuracy was 74.34%. For topics the accuracy was 76.28%. The overall accuracy is of the dimension classifier when the manually constructed map was used was 71.68%. The accuracy if we consider a misclassification only when both classifiers are wrong is 80.42%.

When we incorporated all 14 linguistic indicators, we had the problem that there was not enough training data for the machine learning algorithms to learn the concept.

However, we do show that in principle linguistic indicators can help classification by applying this technique to a smaller problem. We restricted the indicators used to just document length and sentence length in order to keep from confusing the machine learning algorithm with too many attributes.

The baseline accuracy of a Naive Bayes classifier over this set was 68.24. The classifier that uses two linguistic indicators and Naive Bayes is 74.26%.

Future work remains on how to more effectively incorporate these linguistic indicators into the classification in order to be able to use all meaningful indicators for classification.

Analysis of Results

In the case of category dimensions, there are two types of misclassification. There are the misclassification which misclassify both dimensions and there are misclassification which misclassify only one of the two dimension. In the case where only one misclassification occurred, the error is not a complete misclassification. In real applications, these type of errors are less costly than complete misclassification in the one dimensional classifier.

For the linguistic indicators, we have shown that the indicators can be used in conjunction with a word based classifier for improved results. However, the incorporation of the linguistic indicators requires more training data in order to effectively train the classifier.

The intuition on why the linguistic indicators improved performance was because the linguistic indicators can help discard a document that otherwise appears to be in a category based on its words. For example, an extremely short document that contains terms which are associated with international news may actually be a travel document and should be classifier in "other". The linguistic indicators give this kind of information to the classifier.

Conclusions: Implications and Generalizations

The linguistic indicators and the dimension decomposition improve the results from the baseline flat classifier. But in addition to simply improving the results of the classifier these techniques have other implications.

Because the linguistic indicators give the classifier information about structure, this increases the ability of the system to identify documents which might otherwise be misclassified. In applications in heterogeneous environments such as the World Wide Web, there are many articles which appear to have the words to be classified in a category, but their structure is different from documents in that category.

The decomposition also serves to extract features from the document. The dimensional classifiers can be viewed as information extraction tools. For example, a classifier can be trained to determine the "reading" level of a document. This in itself can be used to obtain information about a document. Classifiers that can be built that attempt to extract features such as tone, language sophistication, tense, and other observable features of a document. This is related to Biber's work.

Future work in this area involves incorporating those features into the classifications scheme and incorporating a hierarchical structure on some of the features in the classification scheme. (Sahami 97) This will allow more flexible representation schemes. In addition, future work involves identifying specifically what indicators can enhance classification performance.

References

Mitchell, T. *Machine Learning.* McGraw Hill, 1997.

Biber, D. *A Typology of English Texts.* Linguistics 27 (1989).

Larkey, L. and W. Croft. *Combining Classifiers in Text Categorization.* SIGIR-96.

Karlgren, J. and D. Cutting. *Recognizing Text Genres with Simple Metrics Using Discriminant Analysis.* Proceedings of Coling 94, Kyoto.

Kessler, B., G. Nunberg, H. Schutze. *Automatic Detection of Text Genre.* ACL-97.

Craven, M., D. Dipasquo, D. Freitag, A. McCallum, T. Mitchel, K. Nigam, and S. Slattery *Learning to Extract Symbolic Knowledge from the World Wide Web.* Submitted to AAAI-98.

Daphne K. and M. Sahami *Hierarchically Classifying Documents Using Very Few Words.* ICML-97.