DOCUMENT CLASSIFICATION BY MACHINE:Theory and Practice

Louise Guthrie
Elbert Walker
New Mexico State University
Las Cruces, New Mexico 88001
Joe Guthrie
University of Texas at El Paso
El Paso, Texas 79968

Abstract

In this note, we present results concerning the theory and practice of determining for a given document which of several categories it best fits. We describe a mathematical model of classification schemes and the *one* scheme which can be proved optimal among all those based on word frequencies. Finally, we report the results of an experiment which illustrates the efficacy of this classification method.
TOPICAL PAPER
Subject Area: TEXT PROCESSING

1 Introduction

A problem of considerable interest in Computational Linguistics is that of classifying documents via computer processing [Hayes, 1992; Lewis 1992; Walker and Amsler, 1986]. Simply put, it is this: a document is one of several types, and a machine processing of the document is to determine of which type. In this note, we present results concerning the theory and practice of classification schemes based on word frequencies. The theoretical results are about mathematical models of classification schemes, and apply to any document classification problem to the extent that the model represents faithfully that problem. One must choose a model that not only provides a mathematical description of the problem at hand, but one in which the desired calculations can be made. For example, in document classification, it would be nice to be able to calculate the probability that a document on subject i will be classified as on subject i. Further, it would be comforting to know that there is no better scheme than the one being used. Our models have these characteristics. They are simple, the calculations of probabilities of correct document classification are straightforward, and we have proved that there are no schemes using the same information that have better success rates. In an experiment the scheme was used to classify two types of documents, and was found to work very well indeed.

2 The Description of a Classification Scheme

Suppose that we must classify a document into one of k types. These types are *known*. Here, k is any positive integer at least 2, and a typical value might be anywhere from 2 to 10. Denote these types $T_1, T_2, \ldots, T_k$. The set of words in the language is broken into m disjoint subsets $W_1, W_2, \ldots, W_m$. Now from a host of documents, or a large body of literature, on subject T_i, the frequencies p_{ij} of words in W_j are determined. So with subject T_i we have associated the vector of frequencies $(p_{i1}, p_{i2}, \ldots, p_{im})$, and of course $p_{i1}+p_{i2}+\ldots+p_{im} = 1$. Now, given a document on one of the possible k subjects, it is classified as follows. The document has n words in it, n_1 words from W_1, n_2 words from $W_2, \ldots,$ and n_m words from W_m. Based on this information, a calculation is made to determine from which subject the document is most likely to have come, and is so classified. This calculation is key: there are many possible calculations on which a classification can be made, but some are better than others. We will prove that in this situation, there is a best one.

We elaborate on a specific case which seems to hold promise. The idea is that the frequencies $(p_{i1}, p_{i2}, \ldots, p_{im})$ will be different enough from i to i to distinguish between types of documents. From a document of word length n, let n_j be the number of words in W_j. Thus the vector of word frequencies for that particular document is $(n_1/n, n_2/n, \ldots, n_m/n)$. The word frequencies ¿from a document of type i should resemble the frequencies $(p_{i1}, p_{i2}, \ldots, p_{im})$, and indeed, the classification scheme is to declare the document to be of type T_i if its frequencies "most closely resemble" the frequencies $(p_{i1}, p_{i2}, \ldots, p_{im})$. Intuitively, if two of the vectors are $(p_{i1}, p_{i2}, \ldots, p_{im})$ very nearly equal, then it will be difficult to distinguish documents of those two types. Thus the success of classification depends critically on the vectors $(p_{i1}, p_{i2}, \ldots, p_{im})$ of frequencies. Equivalently, the sets W_j are critical, and must be chosen with great care. The particular situation we have in mind is this. Each of the types of documents is

on a rather special topic, calling for a somewhat specialized vocabulary. The language is broken into $k+1$ disjoint sets $W_1, W_2, \ldots, W_{k+1}$ of words. For $i \leq k$, the words in W_i are "specific" to subject i, and W_{k+1} consists of the remaining words in the language. Now from a host of documents, or a large body of literature, on the subject T_i, we determine the frequencies p_{ij} of words in W_j. But first, the word sets W_i are needed, and it is also from such bodies of text that they will be determined. Doing this in a manner that is optimal for our models is a difficult problem, but doing it in such a way that our models are very effective seems quite routine.

So with subject T_i we have associated the vector of frequencies $(p_{i1}, p_{i2}, \ldots, p_{im})$, the vector being of length one more than the number of types of documents. Since the words in W_i are specific to documents of type T_i, these vectors of frequencies should be quite dissimilar and allow a sharp demarkation between document types. This particular scheme has the added advantage that m is small, being $k+1$, only one more than the number of document types. Further, our scheme will involve only a few hundred words, those that appear in $W_1, W_2, \ldots, W_k$, with the remainder appearing in W_{k+1}. This makes is possible to calculate the probabilities of correct classification of documents of each particular type. Such calculations are intractable for large m, even on fast machines. There are classification schemes being used with m in the thousands, making an exact mathematical calculation of probabilities of correct classification next to impossible. But with k and m small, say no more than 10, such calculations are possible.

3 The Mathematical Model

A mathematical description of the situation just described is this. We are given k multinomial populations, with the i-th having frequencies $(p_{i1}, p_{i2}, \ldots, p_{im})$. The i-th population may be envisioned to be an infinite set consisting of m types of elements, with the proportion of type j being p_{ij}. We are given a random sample of size n from one of the populations, and are asked to determine from which of the populations it came. If the sample came from population i, then the probability that it has n_j elements of type j is given by the formula

$$(n!/n_1!n_2!\cdots n_m!)(p_{i1}^{n_1}p_{i2}^{n_2}\cdots p_{im}^{n_m}).$$

This is an elementary probabilistic fact. If a sample to be classified has n_j elements of type j, we simply make this calculation for each i, and judge the sample to be from population i if the largest of the results was for the i-th population. Thus, the sample is judged to be from the i-th population if the probability of getting the particular n_j's that were gotten is the largest for that population.

To determine which of

$$(n!/n_1!n_2!\cdots n_m!)(p_{i1}^{n_1}p_{i2}^{n_2}\cdots p_{im}^{n_m})$$

is the largest, it is only necessary to determine which of the $(p_{i1}^{n_1}p_{i2}^{n_2}\cdots p_{im}^{n_m})$ is largest, and that is an easy machine calculation. All numbers are known beforehand except the n_j's, which are counted from the sample.

Before illustrating success rates with some calculations, some comments on our modeling of this document classification scheme are in order. The i-th multinomial population represents text of type T_i. This text consists of m types of things, namely words from each of the W_j. The frequencies $(p_{i1}, p_{i2}, \ldots, p_{im})$ give the proportion of words from the classes $W_1, W_2, \ldots, W_m$ in text of type T_i. A random sample of size n represents a document of word length n. This last representation is arguable: a document of length n is not a random sample of n words from its type of text. It is a structured sequence of such words. The validity of the model proposed depends on a document reflecting the properties of a random sample in the frequencies of its words of each type. Intuitively, long documents will do that. Short ones may not. The success of any implementation will hinge on the frequencies $(p_{i1}, p_{i2}, \ldots, p_{im})$. These frequencies must differ enough from document type to document type so that documents can be distinguished on the basis of them.

4 Some Calculations

We now illustrate with some calculations for a simple case: there are two kinds of documents, T_1 and T_2, and three kinds of words. We have in mind here that W_1 consists of words specific to documents of type T_1, W_2 specific to T_2, and that W_3 consists of the remaining words in the language. So we have the frequencies (p_{11}, p_{12}, p_{13}) and (p_{21}, p_{22}, p_{23}). Of course $p_{i3} = 1 - p_{i1} - p_{i2}$. Now we are given a document that we know is either of type T_1 or of type T_2, and we must discern which type it is on the basis of its word frequencies. Suppose it has n_j words of type j, $j = 1, 2, 3$. We calculate the numbers

$$t_i = p_{i1}^{n_1}p_{i2}^{n_2}p_{i3}^{n_3}$$

for $i = 1, 2$, and declare the document to be of type T_i if t_i is the larger of the two. Now what is the probability of success? Here is the calculation. If a document of size n is drawn from a trinomial population with parameters (p_{11}, p_{12}, p_{13}), the probability of getting n_1 words of type 1, n_2 words of type 2, and n_3 words of type 3 is

$$(n!/n_1!n_2!n_3!)(p_{11}^{n_1}p_{12}^{n_2}p_{13}^{n_3}).$$

Thus to calculate the probability of classifying successfully a document of type T_1 as being of that type, we must add these expressions over all those triples (n_1, n_2, n_3) for which t_1 is larger than t_2. This is a

fairly easy computation, and we have carried it out for a host of different p's and n's. Table 1 contains results of some of these calculations.

Table 1 gives the probability of classifying a document of type T_1 as of type T_1, and of classifying a document of type T_2 as of type T_2. These probabilities are labeled $Prob(1)$ and $Prob(2)$, respectively. Of course, here we get for free the probability that a document of type T_1 will be classified as of type T_2, namely $1-Prob(1)$. Similarly, $1-Prob(2)$ is the probability that a document of type T_2 will be classified as of type T_1. The p_{ij} are the frequencies of words from W_j for documents of type T_i, and n is the number of words in the document.

Table 1

p_{1j}	.08	.04	.88	
p_{2j}	.03	.06	.91	
n	50	100	200	400
$Prob(1)$	.760	.871	.951	.991
$Prob(2)$	.842	.899	.959	.992

p_{1j}	.10	.03	.87	
p_{2j}	.02	.05	.93	
n	50	100	200	400
$Prob(1)$	.894	.963	.995	.999
$Prob(2)$	.920	.975	.997	.999

p_{1j}	.08	.04	.88	
p_{2j}	.07	.04	.89	
n	50	100	200	400
$Prob(1)$	.575	.553	.595	.638
$Prob(2)$	.533	.598	.617	.658

There are several things worth noting in Table 1. The frequencies used in the table were chosen to illustrate the behavior of the scheme, and not necessarily to reflect document classification reality. However, consider the first set of frequencies (.08, .04, .88) and (.03, .06, .91). This represents a circumstance where documents of type T_1 have eight percent of their words specific to that subject, and four percent specific to the other subject. Documents of type T_2 have six percent of their words specific to its subject, and three percent specific to the other subject. These percentages seem to be easily attainable. Our scheme correctly classifies a document of length 200 and of type T_1 95.1 percent of the time, and a document of length 400 99.1 percent of the time. The last set of frequencies, (.08, .04, .88) and (.07, .04, .89) are almost alike, and as the table shows, do not serve to classify documents correctly with high probability. In general, the probabilities of success are remarkably high, even for relatively small n, and in the experiment reported on in Section 6, it was easy to find word sets with satisfatory frequencies.

It is a fact that the probability of success can be made as close to 1 as desired by taking n large enough, assuming that (p_{11}, p_{12}, p_{13}) is not identical to (p_{21}, p_{22}, p_{23}). However, since for reasonable frequencies, the probabilities of success are high for n just a few hundred, this suggests that long documents would not have to be completely tabulated in order to be classified correctly with high probability. One could just use a random sample of appropriate size from the document.

The following table give some success rates for the case where there are three kinds of documents and four word classes. The rates are surprisingly high.

Table 2

p_{1j}	.05	.03	.02	.90
p_{2j}	.01	.06	.01	.92
p_{3j}	.04	.02	.08	.86
n	50	100	200	400
$Prob(1)$	.703	.871	.966	.997
$Prob(2)$	.884	.938	.985	.999
$Prob(3)$	.826	.922	.981	.998

p_{1j}	.05	.03	.02	.90
p_{2j}	.01	.05	.01	.93
p_{3j}	.03	.02	.05	.90
n	50	100	200	400
$Prob(1)$	.651	.784	.906	.978
$Prob(2)$	.826	.917	.977	.998
$Prob(3)$	.697	.815	.916	.978

5 Theoretical Results

In this section, we prove our optimality result. But first we must give it a precise mathematical formulation. To say that there is no better classification scheme than some given one, we must know not only what "better" means, we must know precisely what a classification scheme is. The setup is as in Section 3. We have k multinomial populations with frequencies $(p_{i1}, p_{i2}, \ldots, p_{im})$, $i = 1, 2, \ldots, k$. We are given a random sample of size n from one of the populations and are forced to assert from which one it came. The information at our disposal, besides the set of frequencies $(p_{i1}, p_{i2}, \ldots, p_{im})$, is, for each j, the number n_j of elements of type j in the sample. So the information ¿from the sample is the tuple $(n_1, n_2, \ldots, n_m)$. Our scheme for specifying from which population it came is to say that it came from population i if $(n!/n_1!n_2!\cdots n_m!)(p_{i1}^{n_1}p_{i2}^{n_2}\cdots p_{im}^{n_m})$ is maximum over the i's. This then, determines which $(n_1, n_2, \ldots, n_m)$ results in which classification. Our scheme partitions the sample space, that is, the set of all the tuples $(n_1, n_2, \ldots, n_m)$, into k pieces,

the i-th piece being those tuples $(n_1, n_2, \ldots, n_m)$ for which $(n!/n_1!n_2!\cdots n_m!)(p_{i1}^{n_1} p^{n_2} \cdots p_{im}^{n_m})$ is maximum. For a given sample (or document) size n, this leads to the definition of a scheme as any partition $\{A_1, A_2, \ldots, A_k\}$ of the set of tuples $(n_1, n_2, \ldots, n_m)$ for which $\sum_i n_i = n$ into k pieces. The procedure then is to classify a sample as coming from the i-th population if the tuple $(n_1, n_2, \ldots, n_m)$ gotten from the sample is in A_i. It doesn't matter how this partition is arrived at. *Our* method is via the probabilities

$$q_i(n_1, n_2, \ldots, n_m) = (n!/n_1!n_2!\cdots n_m!)(p_{i1}^{n_1} p_{i2}^{n_2} \cdots p_{im}^{n_m}).$$

There are many ways we could define optimality. A definition that has particular charm is to define a scheme to be optimal if no other scheme has an higher overall probability of correct classification. But in this setup, we have no way of knowing the overall rate of correct classification because we do not know what proportion of samples come from what populations. So we cannot use that definition. An alternate definition that makes sense is to define a scheme to be optimal if no other scheme has, *for each population*, a higher probability of correct classification of samples from that population. But our scheme is optimal in a much stronger sense. We define a scheme $A_1, A_2, \ldots, A_k$ to be optimal if for any other scheme $B_1, B_2, \ldots, B_k$,

$$\sum_i P(A_i|T_i) \geq \sum_i P(B_i|T_i).$$

Proofs of the theorems in this note will be given elsewhere.

Theorem 1 *Let $T_1, T_2, \ldots, T_k$ be multinomial populations with the $i-th$ population having frequencies $(p_{i1}, p_{i2}, \ldots, p_{im})$. For a random sample of size n from one of these populations, let n_j be the number of elements of type j. Let*

$$q_i(n_1, n_2, \ldots, n_m) = (n!/n_1!n_2!\cdots n_m!)(p_{i1}^{n_1} p_{i2}^{n_2} \cdots p_{im}^{n_m}).$$

Then the partition of the sample space $\{(n_1, n_2, \ldots, n_k) : n_j \geq 0, \sum_j n_j = n\}$ given by

$$A_i = \{(n_1, n_2, \ldots, n_m) : q_i(n_1, n_2, \ldots, n_m)) > q_j(n_1, n_2, \ldots, n_m) \text{ for } i \neq j\}$$

is an optimal scheme for determining from which of the populations a sample of size n came.

An interesting feature of Table 1 is that for all frequencies $Prob(1) + Prob(2)$ is greater for sample size 100 than for sample size 50. This supports our intuition that larger sample sizes should yield better results. This is indeed a fact.

Theorem 2 *The following inequality holds, with equality only in the trivial case that $p_{ik} = p_{jk}$ for all i, j, and k,*

$$\sum_{n+1} \max_i(((n+1)!/(n_1!n_2!\cdots n_m!)p_{i1}^{n_1} p_{i2}^{n_2} \cdots p_{im}^{n_m}) \geq \sum_n \max_i((n!/(n_1!n_2!\cdots n_m!)p_{i1}^{n_1} p_{i2}^{n_2} \cdots p_{im}^{n_m}),$$

where $\sum_{n+1}$ means to sum over those tuples $(n_1, n_2, \ldots, n_m)$ whose sum is $n+1$, and $\sum_n$ means to sum over those tuples $(n_1, n_2, \ldots, n_m)$ whose sum is n.

6 Practical Results

Our theoretical results assure us that documents can be classified correctly if we have appropriate sets of words. We have algorithms which compute the probability of classifying document types correctly given the document size and the probability of some specialized sets of words appearing in the two document types. Tables 1 and 2 show some sample outputs from that program. Intuitively, we need sets of words which appear much more often in one text type than the other, but the words do not need to appear in either text type very often. Below we describe an experiment with two document collections that indicates that appropriate word sets can be chosen easily. Moreover, in our sample experiment, the word sets were chosen automatically and the classification scheme worked perfectly, as predicted by our theoretical results.

Two appropriate collections of text were available at the Computing Research Laboratory. The first was made up of 1000 texts on business (joint ventures) ¿from the DARPA TIPSTER project and the second collection consisted of 1100 texts from the Message Understanding Conference (MUC) [Sundheim, 1991] describing terrorist incidents in South America. The business texts were all newspaper articles, whereas the MUC texts were transmitted by teletype and came from various sources, such as excerpts from newspaper articles, radio reports, or tape recorded messages. The collections were prepared by human analysts who judged the relevance of the documents in the collections. Each collection contained about half a million words.

We removed any dates, annotations, or header information from the documents which uniquely identified it as being of one text type or another. We divided each collection of texts in half to form two training sets and two test sets of documents, yielding four collections of about a quarter of a million words each. We treated each of the training sets as one huge text and obtained frequency counts for each of the words in the text. Words were not stemmed and no stop list was used. The result was two lists of words with their corresponding frequencies, one for the TIPSTER training set and one for the MUC training set.

Our goal at this point was to choose two sets of words, which we call TIP-SET and MUC-SET, that could be used to distinguish the documents. We knew from the results of TABLE 1 that if we could identify one set of words (TIP-SET) that appeared in the TIP-

STER documents with probability .1 and in the MUC documents with low probability (say .03 or less) and another set (MUC-SET) that appeared with probability .1 in the MUC documents and a low probability (say .03 or less) in the TIPSTER documents, that we could achieve perfect or nearly perfect classification. We used a simple heuristic in our initial tests: choose the TIP-SET by choosing words which were among the 300 most frequent in the TIPSTER training set and not in the 500 most frequent in the MUC training set. We intended to vary the 300 and 500 to see if we could choose good sets. However, this algorithm yielded a set of words that appeared with probability .13 in the TIPSTER training set and with probability .01 in the MUC training set. Note that even though no stop list was used when the frequency counts were taken, this procedure effectively creates a stop list automatically. The same algorithm was used to create the MUC-SET: choose words from among the 300 most frequent in the MUC training set if they did not appear in the 500 most frequent in the TIPSTER training set.

Our theoretical results implied that we could classify each document type correctly 99.99% or the time if we had documents with at least 200 words. Our average document size in the two collections was 500 words. We then tested the classification scheme on the remaining half (those not used for training) of each document set. Only one document was classified differently from the human classification.

When we read the text in question, it was our opinion that the original document classification by a human was incorrect. If we change the classification of this text, then our document classification scheme worked perfectly on 700 documents. It should be noted that the two document collections that were available to us were on very different subject matter, so the choice of the word sets was extremely easy. We expect that differentiating texts which are on related subject areas will be much more difficult and we are developing refinements for this task.

7 References

[Hayes, 1992] Philip Hayes, Intelligent High-Volume Text Processing Using Shallow, Domain Specific Techniques, *Text-Based Intelligent Systems*, P. Jacobs, ed., Lawrence Erlbaum, Hillsdale, NJ, pp. 227 - 241.

[Lewis, 1992] David Lewis, Feature Selection and Feature Extraction for Text Categorization, *Proceedings Speech and Natural Language Workshop*, Morgan Kaufman, San Mateo, CA, February 1992, pp. 212 - 217.

[Sundheim, 1991] Beth Sundheim, editor. *Proceedings of the Third Message Understanding Evaluation and Conference*, Morgan Kaufman, Los Altos, CA, May 1991.

[Walker and Amsler, 1986] D. Walker and R. Amsler, The Use of Machine-Readable Dictionaries in Sublanguage Analysis, *Analyzing Language in Restricted Domains*, Grishman and Kittredge, eds., Lawrence Erlbaum, Hillsdale, NJ.