Improving on the Naïve Bayes Document Classifier

yyyyyyyyy@zzzzzzzz

Abstract

The Naïve Bayes document classifier has been used in many document classification algorithms [1], but is only really useful on a small subset of documents due to it’s many shortcomings [2]. By augmenting the basic functionality of the simple Naïve Bayes classifier, the classification algorithm can be applied to a much wider range of documents. This paper investigates the advantages which can be obtained by adding Feature Selection, Binary Independence, and the Multinomial model to the Naïve Bayes classifier.

1 Introduction

After reading a given document, a human, more often than not, is easily able to categorize and/or determine the relevance of the document. The contextual richness of human language makes this a difficult task for computer algorithms to approximate [4]. The human mind is able to look at the document as a whole, drawing observations about one part of a document from other portions of the document and fully understanding the overall purpose of the document. Many classification algorithms, in contrast, are interested only in the terms present in the document, or other features, based on the algorithm [2]. One such classifier is the simple Naïve Bayes classifier. In this paper, I will show how this classifier can be improved by augmenting it’s functionality with other classification techniques.

2. Format

This paper first defines what is meant by a document in section 3. This explanation is used thought the remainder of the paper. Section 4 explores the Naïve Bayes assumption itself and divulges its weaknesses, section 5 defines multiple feature selection techniques and explains why some are preferred over others, section 6 introduces the Binary Independence Model, and section 7 illustrates the use of Poissons and the Multinomial approach to incorporate term frequency into classification.

3. Definition of a Document

In order to fully understand the classification techniques presented, the concept of a document must be examined. There are two separate document constructs which will be of importance in this paper. The first is the representation of a document as a vector of binary attributes representing whether or not a word is contained within the given document [1, 2]. The second is the representation of a document as multiple word events drawn from a given vocabulary [2]. In the first representation, the document is viewed as an individual event, independent of all other document events, while in the second representation, each word is viewed as an event, independent of all other word events.
4. The Simple Naïve Bayes Classifier

The simple Naïve Bayes classification technique uses the first definition of a document. Given a document vector \(d_i\), one can classify a document based on the probability that each element in vector \(d_i\) will appear given class \(c_j\), using the Naïve Bayes assumption \([2, 6]\). Naïve Bayes estimates the probability of a classification given a document using the following equation:

\[
P(c_j | d_i) = \frac{P(c_j) P(d_i | c_j)}{P(d_i)}\]

[6]

equation 1: Naïve Bayes classification

where \(c_j\) is the classification being considered and \(d_i\) is the document in question. Using the conditional independence assumption, \(P(d_i | c_j)\) can be estimated as:

\[
P(d_i | c_j) = \prod_{t=1}^{V} P(w_t | c_j)\]

[6]

equation 2: conditional independence assumption

where \(V\) represents the vocabulary space and \(w_t\) is an individual term within the document vector. This approach works well on documents with a fixed number of attributes \([2]\). When the document space is expanded to include documents of varying size, the simple Naïve Bayes approach runs up against multiple obstacles.

The Naïve Bayes approach, and the representation of a document as a vector of binary attributes, inherently ignores document length and term frequency \([1]\). This is a critical problem since a longer document is more likely to contain many low frequency words which will greatly skew the classification of the document. Thus, the Naïve Bayes approach will favor long documents since said documents are more likely to contain virtually any word in the vocabulary. Another problem which the Naïve Bayes classifier comes up against is that of the need for the probability of a document class occurring. Such a probability is hard, if not impossible to find for every class of document being investigated \([1]\). Large term vector space also poses a problem as it increases computation time and increases the chance that terms irrelevant to the classification of the document will have large influence over the classification as a whole \([3]\). Interestingly enough, Naïve Bayes independence assumption does not have a large negative effect on the classification of documents even though it violates the natural understanding that terms in a document are indeed related to other terms in the document. This is, because of the fact that classification estimation is a function of the sign of the estimation instead of the value. Thus, it has been shown that approximation of the function can be poor while the classification accuracy remains high \([2]\).
5. Feature Selection

The complication due to the size of the term vector space can be minimized by a technique called Feature Selection. By using feature selection, the algorithm is able to choose a subset of the vocabulary which it finds to be most pertinent to the classification of a set of documents instead of classifying based on the entire vocabulary [2]. This not only decreases the size of the term vector space, it also decreases and/or eliminates the number of noise words which will appear within a document. A noise word is a word which appears (usually at a low frequency) within a document, which has little or no bearing on the classification of the document. Stop words such as “and”, “but”, and “or” are considered noise words and are usually eliminated automatically during document classification [3]. Depending on the class of a document, other words may be deemed as noise words and will be eliminated during feature selection. There are a variety of feature selection techniques, four of which I highlight here:

- **Document Frequency**: The number of documents in which a term occurs. A term is considered important if its frequency is greater than a predetermined threshold.
- **Information Gain**: Measure of the number of bits of information gained by knowing the presence or absence of a term in a document.
- **Mutual Information**: A measure of the association between a word and a document class, based on the number of times a word and a class co-occur, and the number of times each occurs individually.
- **Term Strength**: A value based on the probability that a term will occur within related documents. [3]

All of these techniques will pick and choose words, or terms, from the full vocabulary which are deemed most important for the purposes of classification, thus reducing the term vector space (in some cases up to 98% without loss of classification accuracy [3]).

Research conducted at Carnegie Mellon University by Yiming Yang [3] brought to light a conclusion which greatly aids further enhancement to the Naïve Bayes classifier, as well as guides the choice of a feature selection algorithm. Based on Yiming Yang’s research, the Document Frequency and Information Gain algorithms greatly outperformed the Mutual Information and Term Strength algorithms. This observation led to further investigation concerning the similarities of the various algorithms. It was found that both Document Frequency and Information Gain greatly favor common words in documents, a common word being a word which appears at a high frequency within a document or class of documents. This finding is in stark contrast to the belief that common terms are less informative than less common keyword terms [3].

6. Binary Independence Model

Using the Binary Independence Model, or BIM, we are able to calculate the probability of a document being within a certain class without needing the probability of the document itself [1]. The BIM assumes that there are only two classes into which a document will fall, which is the case in many text classification problems such as the decision of whether or not a document is relevant. Since there are only two document classes, the following equality holds true:
\[ P(\text{c2} | x) = 1 - P(\text{c1} | x) \]  
\[ \text{equation 3: two class assumption} \]

Given this equality, it has been shown that the probability equation derived from the Naïve Bayes assumption can be manipulated in such a way that the probability of a given class need not be estimated in order to calculate an estimate of the probability of a document class given a particular term vector [1]. Adding BIM to the Naïve Bayes classifier allows the algorithm to search through an infinite number of documents without needing to calculate the probability of each classification.

### 7. Compensating for Word Frequency and Document Length

The BIM still falls short of the goal though as it can not account for word frequency or document length. In fact, given the representation of a document as a term vector with binary values, it would be quite difficult to account for frequency of terms and document length. However, by using the second representation of a document, a document as multiple word events drawn from a given vocabulary, these factors can be accounted for [2].

#### 7.1 Poisson approach

Attempts have been made to overcome the shortcomings of the BIM model by incorporating various Poisson mixtures [1]. By using a Poisson, the algorithm can calculate the probability that an event (word) will occur a certain number of times within a document [5]. This differs from previous methods in the sense that previous methods based class on the presence or absence of a term, while the Poisson bases class on the frequency of a term.

In order to use the Poisson approach, the collection of documents must be viewed as a Poisson Process, and thus must satisfy the following guidelines:

1. The number of equal words within nonoverlapping intervals is independent for all intervals
2. The probability of exactly one word occurring in a sufficiently small interval is equal to the probability of the word occurring divided by the number of trials.
3. The probability of two or more changes in a sufficiently small interval is essentially 0. [5]

Thus, this method will only produce the desired results if it is presented with a very large number of documents, as the number of intervals must approach infinity in order for all of the above conditions to hold.

Unfortunately, most Poisson mixtures have not been shown to produce better results than the BIM itself [1]. The failure of the Poisson is based on many factors, including the large number of parameters which must be estimated to calculate the Poisson and an often poor fit to real world data [1].

#### 7.2 Multinomial approach

Another technique which can be used to accommodate for document length and term frequency is what is known as a Multinomial approach. This approach is appealing as it gives high classification power to common terms, as was shown to be useful in section 5. The Multinomial
approach is based on a Multinomial distribution of events (words) where each event is mutually independent with a given probability. Given this, the probability that an event occurs a certain number of times within a given context can be found [5]. In terms of document classification this means that the probability of a document is not simply based on the presence or absence of a word, but on the frequency of the word. This sounds very similar to the description of the Poisson model, but differs in the computational complexity and in the fact that the Multinomial model does not require the collection of documents to fit the parameters of a Poisson Process.

Using the Multinomial model, we can obtain the probability that a document was created by a certain class based on the following equation:

\[
P(d_i | c_j) = \frac{\prod_{t=1}^{V} P(w_t | c_j)^{N_{it}}}{N_i!}
\]

where \(d_i\) is the document in question, \(c_j\) is the class in consideration, \(V\) is the vocabulary space, \(w_t\) is a word or term event within the document, and \(N_{it}\) is the number of times word \(w_t\) occurs within document \(d_i\). Revisiting equations (1) and (2), the conditional independence assumption which was used to estimate \(P(d_i | c_j)\) can be replaced by the Multinomial model equation illustrated in equation (4).

It is clear then that the Multinomial approach gives high probability to words which appear more frequently within a document. As an example, let us say that the term ingredients has a high probability of occurring within documents classified as recipes. Given two documents, \(d_1\) and \(d_2\), where \(d_1\) contains 2 occurrences of the term ingredients, and \(d_2\) contains 30 occurrences of the term, one can see, based on the above equation, that the probability of \(d_2\) having been created by the recipe class is high while the probability that \(d_1\) was created by said class is low, based on the frequency of the term ingredients (note, there may be other terms in document \(d_1\) which raise the probability of the document having been created by the recipe class, but for the purposes of this example, other terms will be ignored). The Multinomial model also allows for increased classification accuracy over documents of varying length, as longer documents will no longer be favored since low frequency occurrences of words will not contribute much to the overall probability estimation.

8. Conclusion

Given a fixed document size, the simple Naive Bayes technique of document classification suffices in most cases. However, real world situations often demand that a document classification algorithm be able to handle documents of wildly varying lengths and varying vocabulary complexities. While the Naive Bayes classifier by itself is not suited for such tasks [2], it can be augmented so as to overcome many obstacles of document classification. By utilizing feature selection based on common word occurrences, the classifier can minimize problems caused by large and complex vocabularies, and can shorten long documents, often containing many noisy terms, down to a manageable collection of important terms. Adding the Binary Independence Model onto this allows the classifier to make estimations of the class of a document without needing an estimation of the probability of the class itself. Finally, utilizing a
technique such as the Multinomial model allows for probability estimation based on term frequency, thus giving higher importance to terms of higher frequency and compensating for document length.

References

   - http://citeseer.ist.psu.edu/lewis98naive.html

   - http://citeseer.ist.psu.edu/mccallum98comparison.html

   - http://citeseer.ist.psu.edu/yang97comparative.html

