PROBABILISTIC TOKEN SELECTION VIA FISHER'S METHOD IN TEXT CLASSIFICATION

Authors: Anqi Fu, Yiming Sun, Katherine Yu
{anqif, sunstat, yukather}@stanford.edu

Abstract. In this project we consider a multiclass text classification problem on three newsgroups with 1,000 entries each with a feature class consisting of over 50,000 tokens. Our baseline Naive Bayes method gives a misclassification error rate of 4.51%, and we focus on variable selection methods to improve upon this error. We compare a token selection method using Naive Bayes to one using the related Fisher's method and a threshold. We find that token selection with Fisher's method does significantly better using two effective choices for this threshold: first, one that controls the number of tokens selected to be 1/5 of the original 53,975; second, a threshold we compute based on a probabilistic estimate of the distribution of the test statistic. The final misclassification errors are 2.60% and 2.34%, respectively.

1. Introduction

In this project, we explore the well-known multiclass text classification using entries from three newsgroups titled *comp.os.ms-windows.misc*, *soc.religion.christian*, and *talk.politics.guns* taken from the 20 Newsgroups dataset [1]. Our data consists of 1,000 entries for each of the three groups, with a feature set of frequencies of 53,975 distinct total tokens. Naturally, our primary objective is to reduce the dimension through variable selection. We observe first that token selection through Naive Bayes does much better than variable selection through PCA, as is often the case with sparse, discrete data. We then improve upon this baseline Naive Bayes by introducing a related technique, Fisher's method, which estimates probabilities of categories given tokens, and which allows us more flexibility and control over the number of selected tokens.

Using Fisher's method, we implement a threshold to select most-distinguishing tokens and de-noise the non-important tokens. We first try cross-validation to choose this threshold, but ultimately, we find two more effective methods to choose the threshold: one, by selecting the threshold that reduces the number of tokens by a factor of 1/5, and two, by computing a threshold based on a probabilistic estimate of the distribution of our test statistic. Using Fisher's method with these two thresholds, we reduce our baseline Naive Bayes misclassification error of 4.51% to 2.60% and 2.34%, respectively.

We will first present the theory of Fisher's method and then show how we used it to improve upon Naive Bayes.

2. Fisher's Method

We understand Fisher's method in relation to Naive Bayes. In Naive Bayes, we model the frequency of each token given the category as a multinomial distribution, and under this model, we estimate the probability of the instance of token j given category k using Laplace smoothing as

$$\phi_{j|y=k} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n_i} 1\{x_j^{(i)} = K \wedge y^{(i)} = k\} + 1}{1\{y^{(i)} = k\}n_i + |V|}, \tag{2.1}$$

where m is the number of documents, n_i is the number of tokens in document i, and $|V|$ is number of tokens in our feature set. In Fisher's method, we instead look at $\mathbb{P}(y = k \mid x = j)$, which is the probability that the document is in category k given that it contains token j. Our test statistic is then

$$T_k = -2\sum_{j=1}^{n} \log P_{Y=k|X=j}. \tag{2.2}$$

This statistic comes from the chi-square statistic of the standard statistical Fisher's method, which is based on the fact that if we simultaneously test n hypotheses, their p-values under the null hypothesis will be distributed as $Unif(0,1)$, so under this assumption, the statistic T_k will be distributed as $\chi^2_{(2n)}$. Lower p-values, which indicate the null is unlikely, lead to higher values of T_k.

The test statistic in our model has a slightly different null distribution. Under our setting, our null hypothesis for each token is that the token occurrence is independent of categories, i.e. given that we have V_j total occurrences of token j over all categories, these occurrences are equally distributed across categories. Our estimate of this probability is

$$P_{Y=k|X=j} = \frac{w_{k,j}}{\sum_{k=1}^{G} w_{k,j}},$$

where G is the number of categories, $w_{i,k} = \phi_{j|Y=k}$, and $\phi_{j|Y=k}$ comes from our Naive Bayes computation. Under the null hypothesis for token j,

$$w_{k,j} \mid \sum_{k=1}^{G} w_{k,j} \sim \text{Mult}\left(\sum_{k=1}^{G} w_{k,j}, 1/G\right), k = 1, ..., G,$$

in other words, $w_{k,j} \mid \sum_{k=1}^{G} w_{k,j} = \sum_{i=1}^{V_j} X_i$, where $X_i \sim Mult(1, 1/G)$ and $V_j = \sum_{k=1}^{G} w_{k,j}$. So when V_j is large, we can use the Central Limit Theorem to approximate this distribution

$$\sqrt{V_j} \begin{pmatrix} P_{1j} \\ \vdots \\ P_{Gj} \end{pmatrix} \xrightarrow{\mathcal{D}} \mathcal{N}\left(\begin{pmatrix} \frac{1}{G} \\ \vdots \\ \frac{1}{G} \end{pmatrix}, \begin{pmatrix} \frac{1}{G}(1-\frac{1}{G}) & & \frac{1}{G^2} \\ & \ddots & \\ \frac{1}{G^2} & & \frac{1}{G}(1-\frac{1}{G}) \end{pmatrix}\right), \tag{2.3}$$

as $V_j \to \infty$.

The assumption of V_j large is reasonable for the high-frequency tokens we will be selecting, thus we assume the test statistic in (2.2) will behave as a non-central log-normal distribution under the null hypothesis. The statistic $-2 \log P_{Y=k|X=j}$ represents a propensity score toward categories for a specific feature; the more likely the category k, the smaller this statistic. The statistic T_k "averages" these propensity scores over the tokens in our feature set. From this null distribution (2.3), we derive an approximate quantile for T_k, which is the probability that T_k would take on such a high value if instances of tokens were independent of categories, and we classify the document in the category k which maximizes this quantile.

3. Variable Selection Methods

Our baseline Naive Bayes algorithm on the entire dataset (with all tokens) gives a classification error of 4.51%. We aim to reduce this error with effective variable selection methods.

We use variable selection through PCA as a reference point. Figure (1) shows the results of the test and train error of multinomial logistic regression and KNN using variable selection through PCA. The best achievable test error rate in logistic regression is .120, which is even bigger than our baseline Naive Bayes.

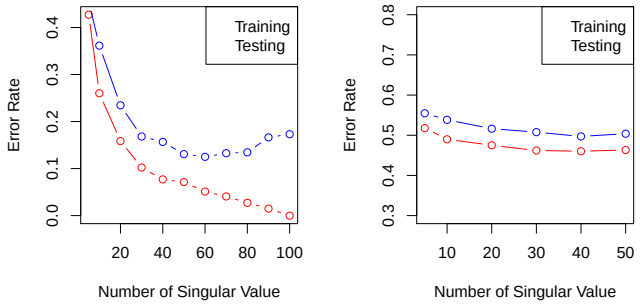

Figure 1. Multinomial Logistic Regression and KNN with variable selection through PCA.

Thus, we dispense with PCA. PCA is a very general method which distorts the specific nature of our data: our frequency table is discrete, while PCA forces us to treat the data as continuous. Also, finding the SVD is unnecessarily time-consuming, especially for a large, sparse matrix. We find token selection much more appropriate and effective on our data.

3.1. **Selecting Most-Distinguishing Tokens.** For comparison with Fisher's method, in Naive Bayes, we select the tokens j for which the variance

$$S^2(\phi_j) = \frac{1}{G-1}\sum_{k=1}^{G}(\phi_{j|Y=k} - \bar{\phi}_j)^2,$$

is maximal, where $\bar{\phi}_j = \frac{1}{G}\sum_{k=1}^{G}\phi_{j|Y=k}$. This is just one measure of how much token j distinguishes categories. Choosing up to 10,000 tokens in this way, the best achievable test error with Naive Bayes is 3.47%.

In Fisher's method, for each token j, we use the statistic

$$M_j = \frac{\max_{k=1,\cdots,G} w_{j,k} - \frac{1}{G}}{\min_{k=1,\cdots,G} w_{j,k} - \frac{1}{G}} \tag{3.1}$$

to determine most-distinguishing tokens. Note that we must have $M_j < 0$. Given that we will later consider the distribution of this statistic under the null hypothesis, instead of taking the N tokens with the lowest M_j,

$$\text{we choose all tokens } j \text{ such that } M_j \leq r_0,$$

where r_0 is a chosen threshold.

In the same feature set size range, the best achievable test error is 2.17%. We want to find an automated method to select a threshold which gives an error close to this optimal error. In the graph below, note that we are not claiming Fisher's method works better than Naive Bayes as a classification algorithm, but instead that Fisher's method allows us to use a better token selection method (selection through the threshold).

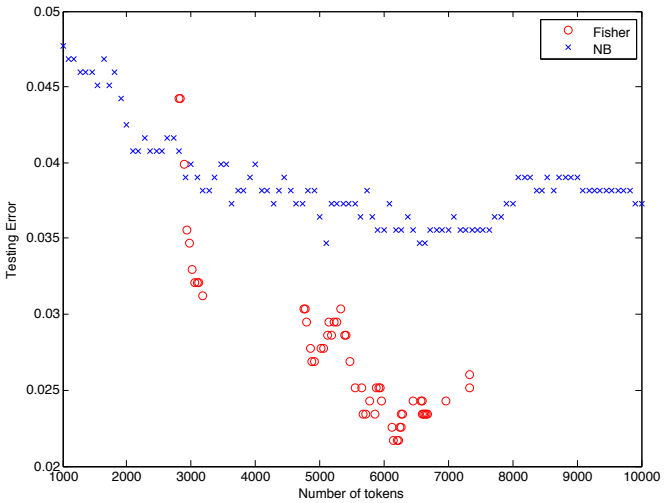

FIGURE 2. Naive Bayes vs. Fisher's method test error with token selection, threshold selection, respectively.

3.2. **Choosing r_0 such that $|V|$ is small.** To select a threshold, we use 10-fold cross-validation, but we do not simply take the threshold which gives minimal average cross-validation error. With our data, many thresholds corresponding to different sizes give about the same error, and we would like to select the smallest one, given the high danger of overfitting. If we simply take the best threshold by cross-validation without restricting size, we get a threshold $r_0 = -1.0447$ corresponding to $42,102$ tokens (85.5% of the total), and a

test error of 3.04%. We find it is very effective to choose the best r_0 that selects at most a fixed fraction of the number of tokens; we restrict to 1/5 of the total number of tokens. This method gives $r_0 = -1.627$ and a test error of 2.6%.

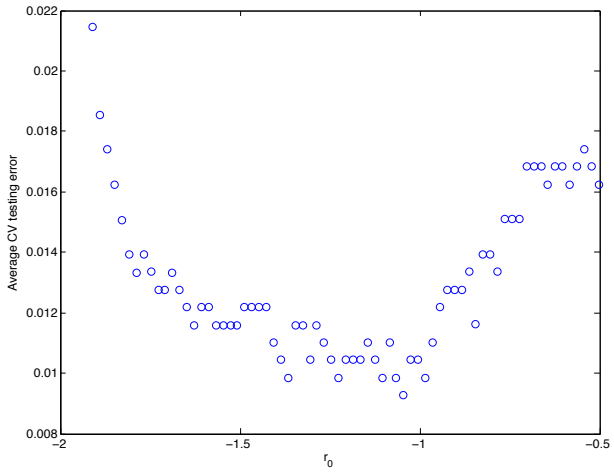

FIGURE 3. 10-fold cross-validation on threshold in Fisher's method.

3.3. **Choosing r_0 based on CLT.** Although restricting size improves on the results of cross-validation, it is still better in some ways to use statistical insights instead. There are many well known issues with CV: in some sense, it is finding the expected minimum value of error, not the parameter which minimizes the expected value of error. Instead, we try to derive a threshold for the statistic M_j in (3.1) based on the asymptotic distribution of $w_{j,k}$ under the null hypothesis. Under the null hypothesis, we have that M_j is asymptotically the ratio of a maximum over a minimum of a G-variate normal distribution with mean 0 and covariance matrix given in (2.3).

We set the threshold r_0 as a quantile for this distribution $\mathbb{P}(M_j < r_0) = 1 - \alpha$ where α is our approximate "size." We use $\alpha = .15$. The method computes this quantile by simulating the statistic M_j under the null hypothesis with the multivariate normal distribution given in (2.3).

The resulting choice of r_0 selects a feature set of 36,133 tokens (73.3 % of the total) and achieves a final test error of 2.34%.

4. Conclusion and future work

Though the feature set selected in (3.3) is very large compared to the size in (3.2), we believe the choice of threshold used in (3.3) is more stable and robust. Both choices are viable for this problem. The choice in (3.2) emphasizes the effectiveness of putting strict limits on the number of tokens (variables) selected to protect against overfitting, but more importantly, we hope the choice of threshold based on the CLT in (3.3) illustrates the value of probabilistic-based methods—how probabilistic interpretations can inform and guide our method choices.

One related idea that we have not explored is that the estimation of quantiles in Fisher's method gives us the flexibility to consider types of classification other than one-to-one classification. For example, we could classify documents into all categories k with quantile for T_k less than 0.10. This method would give classifications with a certain degree of confidence, and would leave others unclassified that we are not confident about. It would also classify documents into multiple categories, which is applicable in many situations, since sometimes, texts or articles can truly belong to more than one topic. Also, in many archiving tasks, it is better to overclassify to make sure each reference text will be located.

References

[1] Lang, Ken. (1995). *20 Newsgroups data set* [Data files]. Retrieved from http://http://qwone.com/ jason/20Newsgroups/
[2] Segaran, Toby. (2008). *Programming Collective Intelligence* [Kindle DX version]. Retrieved from Amazon.com