ISSN: 0976-0172

Journal of Bioscience And Technology
www.jbstonline.com

Text Classification Through the Integration of Probabilistic Classifiers

***Dr.J Nelson*[1]*., G.Jyothi Muthya Sri*[2]**

1 Professor, Department of CSE, Malla Reddy College of Engineering for Women., Maisammaguda., Medchal., TS, India
2, B.Tech CSE (20RG1A05L4),
Malla Reddy College of Engineering for Women., Maisammaguda., Medchal., TS, India

Article Info

Received: 30-06-2023 Revised: 18-08-2023 Accepted: 29-08-2023

Abstract

In the realm of machine learning, probabilistic models are widely regarded as some of the best available. Very little research has been done to evaluate the performance of two or more classifiers used in conjunction in the same classification task, despite the fact that famous probability classifiers show very excellent performance when used separately in a particular classification task. In this study, we employ two probability strategies for document classification: the naïve Bayes classifier and the Maximum Entropy model. Then, we merge the two sets of findings using two different operators—Max and Harmonic Mean—to boost the categorization performance. Results from an evaluation conducted on the "ModApte" subset of the Reuters-21578 dataset demonstrate that the suggested technique improves final evaluation accuracy significantly.

Introduction

One way to look at text categorization is as the process of using a learning model to determine the broad groups that best describe a given set of documents. Each new paper is run through this algorithm, and then placed in a category or multiple categories. Spam screening, e-mail forwarding, online database upkeep, and news filtration are just some of the many uses for text categorization. Over the years, researchers have focused on improving text categorization by studying subjects like effective training and application, adjusting performance, and creating comprehensible models. Multivariate regression models, closest neighbour classifications, stochastic Bayesian models, decision trees, neural networks, and many more statistical classification and machine learning methods have been applied to text segmentation. (Dumais et al., 1998).

Text categorization using SVMs has been investigated previously (Dumais et al., 1998). (Galathiya, 2012). Discriminative methods like Logistic Regression (LR) and Support Vector Machines (SVMs) are one group, while probabilistic approaches such as the aspect model (Hofmann, 1999), the maximum entropy model (Fragoso et al.), the latent Dirichlet allocation (Blei et al., 2002), and the Bayesian classification (Hamad, 2007) are another. (Grossman et al, 2005). Very little research has been done to evaluate the performance of two or more classifiers when used in conjunction in the same classification task, despite the fact that famous classifiers show very excellent performance when used separately in a particular classification task. In this study, we use the naïve Bayes classifier and the Maximum Entropy classification model, two probability methods, to categorize texts. We suggest two combining operators—Max and Harmonic Mean—to join the outputs of the two algorithms in order to boost categorization performance.

Statistical Methods for Classifying Documents

Intuitive Bayesian classification

A text classification can be thought of as a function that converts a document's word count from x to a probability that it corresponds to a certain text group, where x is the number of words in the document. The Naive Bayes classifier (Al-Aidaroos et al, 2010) is commonly used to predict the likelihood of each category if the characteristics x1,xn are conditionally independent, given the category variable c. It is possible to make educated guesses using the Bayes theorem:

$$\Pr(c \mid d) = \frac{\Pr(d \mid c)P(c)}{P(d)}$$

To determine the optimal class (argmaxc Pr(c)Pr(d|c)) for the test set documents, Fragos et al. (2005) utilized training data to predict model parameters. McCallum and Nigam's method served as inspiration for this one. (1998).

Categorization Based on Maximum Entropy

Shannon's theory of transmission makes use of entropy. In and of itself, entropy H quantifies the typical degree of ambiguity associated with a singular random variable X:

$$H(p) = H(X) = -\sum_{x \in X} p(x)\log_2 p(x)$$

where the probability mass function of X is denoted by p(x). Entropy has also been applied to problems in natural language processing, among other areas. The exponential form of the unique distribution with maximal entropy has been demonstrated by Della et al. (Della P. et al., 1997). The iterative scaling (IIS) method was used by Fragoso et al. (2005) to classify texts; it is a hill-climbing algorithm for finding the parameters of the maximum entropy model. In Section 3, we discuss the utility of the chi-square goodness-of-fit test as a supplementary connectedness metric in text categorization. The efficacy of classification is discussed further in Section 4, which details the use of two combining operators of the classification findings. The experimental data and assessment outcomes are presented and discussed in Section 5. Section 6 concludes with our findings and offers suggestions for moving forward.

Choosing Features Via the X-Square Test

In the past, the chi-square test was used to pick features for text categorization. The chi-square and information gain were determined to be the most effective measures of word selection by Yang and Pedersen (Yang and Pedersen, 1997). Using weights for selecting model features and assessing each feature's significance in the classification task, Fragoso et al. (Fragoso, 2005) suggested a novel approach to applying Maximum Entropy modelling for text categorization. They used X-square values to assign relative significance to the model's characteristics in place of the traditional Maximum Entropy approach. They put their approach to a series of categorization tests using the Reuters-21578 dataset. Example Using the Reuters-21578 'ModApte' split training dataset, with the categories c1='Acq' and c2'Acq, we wish to determine whether or not the word 'usa' is an effective feature for the categorization in the category 'Acq'. After eliminating all of the stop words, we find that "usa" occurs 1,238 times in the group c1='Acq' and 4,464 times in the other categories (c2'Acq'). There are 125,907 phrases (words) in the 'Acq' class and 664,241 in all the other groups combined. 790,148 words is the grand total. (words). The absence of a relationship between the term "usa" and the class designation "Acq" is the null hypothesis. The probabilities can be calculated.

w='usa' and c1='Acq': E_{11}= (5,702x125,907)/790,148=908.59
w='usa' and c1≠'Acq': E_{12}= (5,702x664,241)/790,148=4,793.4
w≠'usa' and c1='Acq': E_{21}= (784,446x125,907)/790,148=124,998.4
w'≠usa' and c1≠'Acq': E_{22}= (784,446x664,241)/790,148=659,447.6
Then we calculate the X^2 value:

$$X^2 = (1{,}238-908.59)^2/908.59 + (4{,}464-4793.4)^2/4793.4 + (124{,}669-124{,}998.4)^2/124{,}998.4 + (659{,}777-659{,}447.6)^2/659{,}447.6 = 143.096.$$

With one degree of freedom and a significance level of 0.05, we can deny the null hypothesis if the computed value from the X2 distribution table is larger than the crucial value. Therefore, the word "usa" is a suitable feature for the categorization in the category "Acq" if the computed X2 value is big, and we have solid proof for the combination ('usa,' 'Acq'). Combining the results of the Naive Bayes and Maximum Entropy classifications with a "Merging Operator" To account for differences in accuracy between the NBC and MEC, we blend their output using two operators to produce more accurate classifications.

$$MaxC(d) = Max\ \{\mathrm{NBC}(d), \mathrm{MEC}\ (d)\}$$
$$HarmonicC\ (d) = 2.0 \times \mathrm{NBC}(d) \times \mathrm{MEC}\ (d)\ /\ (\mathrm{NBC}(d) + \mathrm{MEC}\ (d))$$

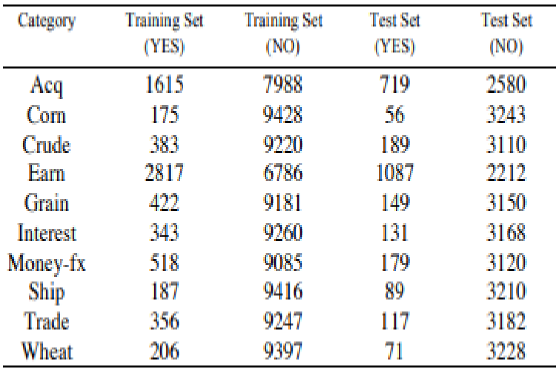
Based on the findings of the Naive Bayes (NBC (d)) and Maximum Entropy classifications (MEC(d)), the MaxC(d) algorithm selects the highest possible value for the incoming document d, as shown in equation (3). The Harmonic Mean of the values from these two classifications is estimated by the Harmonic (d) algorithm in

Equation 4. Online biological papers having Databank Accession Numbers (a vital component of bibliographic material) were classified using these combining operators by Jong woo, Daniel, and George (Jong woo et al, 2010).

Evaluation

The "ModApte" subset of the Reuters-21578 dataset was used to test the suggested categorization method. There are 9,603 sample papers used for training and 3,299 used for testing in the collection. From a pool of 135 options, 10 were ultimately selected. (see table 1). Documents are filed under the "Yes" category if they fit the criteria for that group, and under the "No" category otherwise. Table 1 displays the total number of papers in each of the 10 groups used during the training and testing phases.

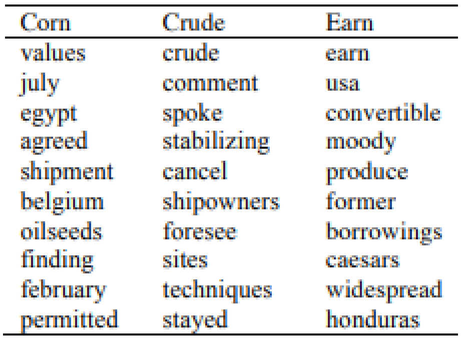
Table 1. 10 categories from the “ModApte” split of the Reuters-21578 dataset with the number of documents for the Training phase and the Test phase



Category	Training Set (YES)	Training Set (NO)	Test Set (YES)	Test Set (NO)
Acq	1615	7988	719	2580
Corn	175	9428	56	3243
Crude	383	9220	189	3110
Earn	2817	6786	1087	2212
Grain	422	9181	149	3150
Interest	343	9260	131	3168
Money-fx	518	9085	179	3120
Ship	187	9416	89	3210
Trade	356	9247	117	3182
Wheat	206	9397	71	3228

All texts were processed using a collection of stop words in both the training and testing phases. Finally, out of a total of 790,148 words, a collection of 32,412 distinct words-terms was established. The maximum entropy model was then applied to the data, and the 2,000 highest-ranked terms in each group were chosen. The X square test's top 10-word phrases across three groups are listed in Table 2.

Table 2. 10 top ranked words calculated by the X square test for three categories of the ModApte Reuters-21578 training dataset



Corn	Crude	Earn
values	crude	earn
july	comment	usa
egypt	spoke	convertible
agreed	stabilizing	moody
shipment	cancel	produce
belgium	shipowners	former
oilseeds	foresee	borrowings
finding	sites	caesars
february	techniques	widespread
permitted	stayed	honduras

The top 2000 keywords in each group were used to create the maximum entropy model's features. We used micro-Recall (Re), micro-Precision (Pr), and a micro-averaged F1 metric to assess the algorithms' success. (micro-F1). Let's say that a represents the total number of documents that the system properly categorized into the class category, b represents the total number of documents that were classed into the class, and d represents the total number of documents that actually pertain to the class. We characterize Pr and Re as

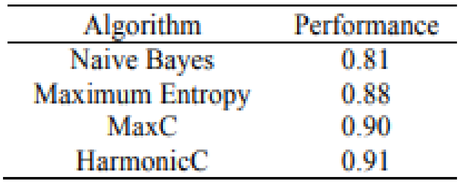
$$\mu \Pr = \frac{\sum_c a}{\sum_c b} \quad \text{and} \quad \mu \mathrm{Re} = \frac{\sum_c a}{\sum_c d}$$

where the summing is over all the classes. The micro-F1 measure is then computed as the harmonic mean of μPr and μRe

$$micro - F1 = 2 \times \mu \Pr \times \mu \mathrm{Re} / (\mu \Pr + \mu \mathrm{Re})$$

Table 3 shows the micro averaged F1 performance Micro-averaged F1 measure performance for Naive Bayes and Maximum Entropy Classifiers and our Max and harmonic merging Operators.

Table 3. Micro-averaged F1 measure performance for Naive Bayes and Maximum Entropy Classifiers and Max and harmonic Operators



Algorithm	Performance
Naive Bayes	0.81
Maximum Entropy	0.88
MaxC	0.90
HarmonicC	0.91

Micro-averaged F1 measure score of 0.88 indicates that the Maximum Entropy classification outperforms Naive Bayes. The Micro-averaged F1 performance of the MaxC(x) and Harmonic(x) operators is higher than that of the Naive Bayes and SVM models.

Conclusion

In this article, we detail a method for classifying texts from the "ModApte" subset of the Reuters-21578 dataset by using a combination of models trained with Naive Bayes and Maximum Entropy. Based on the work of Fragoso et al., we employ a chi-square feature selection approach to zero in on the most informative words-features. (Fragoso, 2005). When compared to the Naive Bayes classification, the Maximum Entropy model appears to work more effectively. To boost efficiency, particularly for the Recall rate, the outputs of the Naive Bayes and SVM models are combined using two combining operators. As can be seen in the outcomes for the Micro-averaged F1 measure, the combining operators do enhance efficiency. (0.90, 0.91 for MaxC and HarmonicC operators respectively). We plan to further enhance efficiency in future work by discovering new ways to gather groups of words-features and new combining algorithms.

References

[1] Al-Aidaroos, K.M., A.A. Bakar, A.A., and Othman, Z., 2010. Naive Bayes variants in classifi-cation learning. In Proceedings of the International Conference on Information Retrieval and Knowledge Management, March 17-18, 2010, Shah Alam, Selangor, pp: 276-281.

[2] Blei, D., Ng A., and Jordan, M. 2002. Latent dirichlet allocation. In Proceedings of NIPS 14.

[3] Della P., S., Della P., V. and Lafferty J., 1997. Inducing features of random fields. IEE trans-action on Pattern Analysis and Machine Intelligence, 19(4).

[4] Dumais, T., S., Platt, J., Heckerman, D., and Sahami, M., 1998. Inductive learning algorithms and representations for text categorization. In Proceedings of the Seventh International Confer-ence on Information and Knowledge Management, pages 148-155. ACM Press.

[5] Fragos, K., Maistros, I., Skourlas, C., 2005. A X2-Weighted Maximum Entropy Model for Text Classification. In Proceedings of 2nd International Conference On Natural Language Understanding and Cognitive Science, Miami, Florida: 22-23.

[6] Galathiya, A. S., Ganatra, A., P., and CK Bhensdadia, K., C., 2012 An Improved decision tree induction algorithm with feature selection, cross validation, model complexity & reduced error pruning, IJSCIT march 2012.

[7] Grossman, D., and P. Domingos, P., 2005. Learning Bayesian Network Classifiers by maxi-mizing conditional likelihood. In Proceedings of the twenty-first international conference on Machine learning, 361–368.

[8] ACM Press. Hamad, A., 2007. Weighted Naive Bayesian Classifier. IEEE/ACS International Conference, on Computer Systems and Applications, AICCSA apos;07, Volume 1, Issue 1, Page(s):437 – 441. Hofmann, T., 1999. Probabilistic latent semantic analysis. In Proceedings of UAI.

[9] Jongwoo, K., Daniel X. L., and George, R., T., 2010. Naïve Bayes and SVM Classifiers For Classifying Databank Accession Number Sentences. National Library of medicine, from Online Biomedical Articles.

[10] McCallum A. and Nigam, K., 1998. A comparison of event models for naive Bayes text classi-fication. In AAAI/ICML-98 Workshop on Learning for Text Categorization. Reuters-21578 http://www.daviddlewis.com/resources/testcollections/reuters21578/

[11] Yang, Y. and Pedersen J., 1997. A comparative study on feature selection in text categorization. In Proceedings of the Fourteenth International Conference on Machine Learning (ICML'97) pp 412-420.