Accurate Text Classification via Maximum Entropy Model

Baoping Zou(✉)

State Grid Info-Telecom Great Power Science and Technology Co., Ltd., Beijing, China
hello_grid80@sina.com

Abstract. Text classification and the research of classification algorithms or models play an important part in the research area of big data, which is among the hottest in our daily life contemporarily. The final target of task of text classification is to choose which is the correct class label that a given text input should belong to. In this paper, we try to propose a more accurate text classification approach by making full use of the principle of maximum entropy model. We conduct a series of experiments of our approach based on a real-world text dataset, which can be downloaded for public research use. The experimental results demonstrate that our proposed approach is very efficient for the task of text classification.

Keywords: Text classification · Maximum entropy model · Big data

1 Introduction

Text classification and the research of classification algorithms or models play an important part in the research area of big data, which is among the hottest in our daily life contemporarily. Nowdays text classification is a necessity for everyone because of the very large amount of text documents that we have to cope with every day and its ever going increasing speed. Against this background, there has been several text classification models to be proposed to solve this problem. In general, text classification models can be divided into two, namely topic-based and genre-based classification models. The former topic-based text categorization classifies text documents according to the topics of the text [1]. Texts can be many genres, which can be represented by a set of words with different weights, for example, scientific articles, news reports, and movie reviews, which is familiar to everybody now. While previous works on the latter genre classification found that this task has some different aspects, large or small, from the former topic-based categorization [2].

As we all know, typically most datasets that we have used for topic classification research are collected on purpose from the web sites such as newsgroups, forums, bulletin boards, email lists or broadcast. Apparently, they come in multi-sources, as a result consequently come with different formats, different sets of vocabularies. And even documents of the same genre have different writing styles. That is to say, the data are nine times out of ten heterogeneous.

S. Wang and A. Zhou (Eds.): CollaborateCom 2016, LNICST 201, pp. 569–576, 2017.
DOI: 10.1007/978-3-319-59288-6_56

As we have described before, intuitively the task of text classification is to classify a given document into a predefined category. An example of the predefined category set may be political, entertainment, sports, finances, and science. More formally, if we let letter d_i represent a document of the entire set of documents D and $\{c_1, c_2, \cdots, c_n\}$ represent the set of all categories, then the target of text classification task is to assign one category c_j to a document d_i. As is done in every supervised machine learning task, an initial training dataset is needed to get the parameters of the model. A document may be assigned to more than one category, but in this paper we only consider assigning a single category to each document. For example, a document may belong to entertainment, and at the same time belong to science too. But we should note that the probability that a document belongs to different categories may be always different.

Maximum entropy model is a general technique that always be used to compute the probability distributions from all sorts of data. The overriding inherent principle in maximum entropy is that when nothing is known, the probability distribution (i.e. the values of the probabilities) should be as uniform as possible, which meets the constraints of maximal entropy from which the model has its name. Labeled training data is used efficiently to derive a set of constraints, i.e. model parameters for the model that characterize the class-specific expectations for the wanted probability distributions. Constraints are represented as expected values of features, which can be any real-valued function of an example. The improved iterative scaling algorithm (IIS) is always used to find the maximum entropy distribution that is most consistent with the former given constraints. While in our text classification scenario, we use maximum entropy to estimate the conditional distribution of the class label of a given a document. A document is represented by a set of word count features with different weights. The labeled training data is used to compute the expected values of the word counts based on a class-by-class basis. The former mentioned Improved iterative scaling is used to find a text classifier of an exponential form that is in line with the constraints from the labeled training data.

Our experimental results demonstrate that maximum entropy principle is a technique that warrants further investigation for the task of text classification, and the proposed maximum entropy model is efficient for this work. On one real data set we used, for example, the maximum entropy model reduces the mean classification error by more or less 40% in comparison to the popular naive Bayes. While on another data sets we used, however, the basic maximum entropy model does not perform as well as naive Bayes. Here, there is apparent evidence that basic maximum entropy suffers from overfitting and poor feature selection, which has a bad influence on the accuracy. When a normal prior is added to the basic maximum entropy model, the classification performance is improved apparently in these cases. Overall, the maximum entropy model we used has a better performance than naive Bayes on two of three data sets. Many research direction of the maximum entropy model for further investigation still exist, which may improve performance even further in the future. These works include more efficient and effective feature selection methods, applying bigrams and phrases as features, and adjusting the appropriate prior knowledge based on the sparsity attributes of the used datasets.

This following of this paper proceeds as follows. Section 2 demonstrate the general framework of the maximum entropy model for computing the conditional probability and distributions. Then, the specific application of maximum entropy to the task of text classification is further discussed in Sect. 3. Related works about the task of text classification and how to apply maximum entropy into it are presented in Sect. 3.1. Experimental results on real datasets are presented in Sect. 3.2. Finally, Sect. 4 discusses our plans for future work.

2 Maximum Entropy Model

The motivating idea behind the maximum entropy principle lies in the fact that one should prefer the most uniform model parameters that also meet every given constraints [3], by which the entropy of the model is largest and this principle got its name the maximum entropy principle. For example, let us consider a four-way text classification task where we are told only that on average 30% of documents with the word children in them are in the school class. Intuitively, when we are given a document with children in it, we would say it has a 30% probability of being a school document, and a 23.3% probability for each of the other three classes. If a document does not have children we would compute the uniform class distribution, 35% each. This model is exactly the maximum entropy model that conforms to our known constraints. Computing the model is easy in this example, but when there are many constraints to meet, rigorous techniques are needed to find the only optimal solution.

In its most general formulation, the maximum entropy principle can be used to compute any probability distribution. In this work, we are more interested in text classification; thus we limit our further discussion to learning more accurate conditional distributions from the labeled training data. Specifically, we learn by the labeled training data by the maximum entropy model that the conditional distributions of the class label given a document.

2.1 Constraints and Features

In maximum entropy we use the labeled training data to set constraints on the conditional distribution. Each constraint expresses a characteristic of the training data that should also be demonstrated in the learned distribution and model parameters. We can make any selected real-valued function with two kinds of parameters, namely the document and the class, be a feature $f_i(d, c)$. Maximum entropy gives us the power to restrict the model distribution to have the same expected value for this feature as seen in the training data D, i.e. the prior knowledge and parameters. Thus, we get that the learned conditional distribution and trained parameters $P(c|d)$ must have the following property:

$$\frac{1}{|D|}\sum_{d\in D} f_i(d, c(d)) = \sum_d P(d) \sum_c P(c|d) f_i(d, c). \tag{1}$$

In practice, the document distribution $P(d)$ is unknown by us, and in fact we do not have to be interested in modeling it. Thus, we make use of our training data without the class labels, as an approximation estimation to the document distribution, and write down the following constraint:

$$\frac{1}{|D|}\sum_{d\in D} f_i(d, c(d)) = \frac{1}{|D|}\sum_{d\in D}\sum_{c} f_i(d, c). \tag{2}$$

Thus, when we use the maximum entropy model, the first step is to identify a set of feature functions that will come into role for the task of text classification process. Then, for each selected feature, measure its expected value over the training data and take it as a constraint for the model distribution.

2.2 Parametric Form

When constraints are estimated in this way, it is guaranteed that a unique distribution exists which will deduce the maximum value of the distribution entropy. Moreover, it can be proved that the distribution is always of the exponential form like this:

$$P(c|d) = \frac{1}{Z(d)}\exp\left(\sum_i w_i f_i(d, c)\right), \tag{3}$$

where each $f_i(d, c)$ is a feature, λ_i is a parameter to be estimated and $Z(d)$ is simply the normalizing factor to ensure a proper probability:

$$Z(d) = \sum_c \exp(\sum_i w_i f_i(d, c)). \tag{4}$$

When the constraints are computed from the labeled training data, the solution to the maximum entropy problem is equivalent to the solution to a dual maximum likelihood problem for models of the same exponential form. Additionally, it is guaranteed that the likelihood surface of the objective function is convex, which ensures a single global maximum and no local maxima. This suggests that this problem has a possible approach for finding one and only one maximum entropy solution. The steps are by the following: 1. First guess any initial exponential distribution of the correct form as a starting point; 2. then, perform the hill climbing algorithm or quasi-Newton algorithm in the potential likelihood space; 3. Iterate step 1 and 2 until the solution not changed. As there is no local maxima, the iteration will converge to the only maximum likelihood solution, which will also be the global maximum entropy solution.

Algorithm 1: BFGS algorithm for maximum entropy model learning

Input: feature function $f_1, f_2, \cdots, f_n$; empirical distribution $\tilde{P}(x, y)$,target function $f(w)$,gradient vector $g(w) = \nabla f(w)$,required precision parameter ϵ;
Output: optimal parameter vector w^*;

1 Initialize parameter vector $w^{(0)}$, get B_0 as positive definite matrix, set $k = 0$;
2 compute $g_k = g(w^{(k)})$;
3 **if** $||g_k|| \leq \epsilon$ **then**
4 stop and get $w^* = w^{(k)}$;
5 **else**
6 continue;
7 by equation $B_k p_k = -g_k$, compute p_k;
8 one dimensional searching, get λ_k satisfying

$$f(w^{(k)} + \lambda_k p_k) = min_{\lambda \geq 0} f(w^{(k)} + \lambda p_k)$$

;
9 set $w^{(k+1)} = w^{(k)} + \lambda_k p_k$;
10 compute $g_{k+1} = g(w^{(k+1)})$;
11 **if** $||g_k|| \leq \epsilon$ **then**
12 stop and have $w^* = w^{(k+1)}$;
13 **else**
14

$$B_{k+1} = B_k + \frac{y_k y_k^T}{y_k^T \delta_k} - \frac{B_k \delta_k \delta_k^T B_k}{\delta_k^T B_k \delta_k}$$

, where,

$$y_k = g_{k+1} - g_k, \delta_k = w_{k+1} - w_k$$

;
15 $k = k + 1$, go to Line 7;
16 **final** ;
17 **return** w^*;

2.3 Parameter Learning

In [4], a number of algorithms for computing the parameters of maximum entropy models, which includes gradient ascent, iterative scaling, conjugate gradient, and variable metric methods. Surprisingly, the standardly used iterative scaling algorithms perform quite poorly in comparison to the others in practical scences. While for almost all of the test problems, a quasi-Newton algorithm such as BFGS, outperforms the other candidates for optimizing. As a result, in this paper, we use the quasi-Newton algorithm to train the parameters of the maximum entropy model.

For a maximum entropy model, we have

$$P_w(c|x) = \frac{\exp(\sum_{i=1}^{n} w_i f_i(d, c))}{\sum_{c} \exp(\sum_{i=1}^{n} w_i f_i(d, c))} \tag{5}$$

With simple mathematic manipulations, we get the objective function by the following

$$\min_{w \in R^n} f(w) = \sum_{d} \tilde{P}(d) \log \sum_{c} \exp(\sum_{i=1}^{n} w_i f_i(d,c)) - \sum_{d,c} \tilde{P}(d,c) \sum_{i=1}^{n} w_i f_i(d,c) \quad (6)$$

Then we can get the gradient vector as follows:

$$g(w) = (\frac{\partial f(w)}{\partial w_1}, \frac{\partial f(w)}{\partial w_2}, \cdots, \frac{\partial f(w)}{\partial w_n})^T, \quad (7)$$

where

$$\frac{\partial f(w)}{\partial w_i} = \sum_{d,c} \tilde{P}(d) P_w(c|d) f_i(d,c) - E_{\tilde{P}}(f_i), \ i = 1, 2, \cdots, n \quad (8)$$

The corresponding quasi-Newton algorithm is shown in Algorithm 1.

3 Experiment

In this section, we conduct extensive experiments on a real-world dataset. First we list the baseline models, then we introduce the dataset, at last, we demonstrate the experimental results.

3.1 Baseline Algorithm and Dataset

We compared the maximum entropy model with a number of baseline classification models. The baseline models and their basic information are as follows:

(1) kNN (k-nearest neighbors; here, k = 10). In kNN, an item is classified by a majority vote of its neighbors.
(2) LRC (Logistic Regression Classifier). LRC measures the relationship between a class label and features by estimating probabilities using a logistic function.
(3) NB (Naïve Bayes). NB applies Bayes' theorem by assuming independence among features.
(4) L-SVM (Linear-form support vector machine). L-SVM is a support vector machine with a linear-form kernel function.

For kNN, LRC and NB, we employed scikit-learn, whereas for L-SVM, we chose Weka. All models were used with default settings and parameters. It is worth noting that the implementation of L-SVM in Weka derives from LIBSVM, a well-known library for support vector machines.

We downloaded a real dataset, namely the Reuters-21578 dataset, from David Lewis' page[1]. And we applied the standard train/test split to get the training dataset and

[1] http://www.daviddlewis.com/resources/testcollections/reuters21578/.

test dataset. These documents all firstly appeared on the Reuters newswire in 1987, and then were manually classified by personnel from Reuters Ltd.

For the fact that the class distribution values for these documents is very skewed, two sub-collections are usually taken into account for text categorization tasks.

3.2 Experimental Results

Table 1 shows the comparison results between the maximum entropy model and the baseline models. We can see that the maximum entropy model performs better than the baseline models. The extensive use of many mature models from state-of-the-art machine learning packages as baseline guarantees a comprehensive experimental comparison.

Table 1. Comparison between the maximum entropy model and baseline models.

Models	Accuracy	F1-measure
Maximum Entropy Model	**0.78**	**0.81**
kNN	0.64	0.76
LRC	0.65	0.68
NB	0.69…	0.63
L-SVM	0.73…	0.77

4 Conclusion

In this paper, we investigate how to apply the maximum entropy model into text classification, and find that maximum entropy model give a more accurate classification results than baseline models. Firstly, we try to propose a more accurate text classification approach by making use of maximum entropy model. Then we conduct a series of experiments of our approach based on a real-world text dataset. At last the experimental results demonstrate that our proposed approach is very effective and efficient for text classification task. In the future, we will focus on how to apply the maximum entropy model for QoS measurement [5–7].

References

1. Yang, Y.: An evaluation of statistical approaches to text categorization. Inf. Retrieval **1**(1–2), 69–90 (1999)
2. Kessler, B., Numberg, G., Schütze, H.: Automatic detection of text genre. In: Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics, pp. 32–38. Association for Computational Linguistics (1997)
3. Berger, A.L., Pietra, V.J.D., Pietra, S.A.D.: A maximum entropy approach to natural language processing. Computational Linguistics **22**(1), 39–71 (1996)

4. Malouf, R.: A comparison of algorithms for maximum entropy parameter estimation. In: Proceedings of the 6th Conference on Natural Language Learning, vol. 20, pp. 1–7. Association for Computational Linguistics, August 2002
5. Wang, S.G., Sun, Q.B., Yang, F.C.: Towards web service selection based on QoS estimation. Int. J. Web Grid Serv. **6**(4), 424–443 (2010)
6. Ma, Y., Wang, S., Hung, P.C.K., Hsu, C.-H., Sun, Q., Yang, F.: A highly accurate prediction algorithm for unknown web service QoS value. IEEE Trans. Serv. Comput. **99**, 1–10 (2015). doi:10.1109/TSC.2015.2407877. http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=7051279&tag=1
7. Wang, S., Ma, Y., Cheng, B., Yang, F., Chang, R.N.: Multi-dimensional QoS prediction for service recommendations. IEEE Trans. Serv. Comput. (2016). doi:10.1109/TSC.2016.2584058. https://www.computer.org/csdl/trans/sc/preprint/07498681.pdf