IMPROVED TOPIC CLASSIFICATION OVER MAXIMUM ENTROPY MODEL USING K-NORM BASED NEW OBJECTIVES

Xiang Li, Ea-Ee Jan, Cheng Wu and David Lubensky

Human Language Technology Department
IBM T.J. Watson Research Center
Yorktown Heights, New York, 10598, USA

{xiangli, ejan, chengwu, davidlu}@us.ibm.com

ABSTRACT

Maximum Entropy (MaxEnt) model has been proven to be a very effective approach in the topic classification task, where a specific topic from a pre-defined topic set will be assigned to each sentence. Although it is originally developed based on the motivation of maximizing the conditional probability entropy under certain constraints, MaxEnt model is indeed an exponential distribution model that maximizes the log-likelihood of the training data. This log-likelihood criterion bears similarity with the classification accuracy criterion, which is the ultimate performance measure of a topic classifier. But these two criterion still differ from each other, and their discrepancy consequently reduces the benefit of optimization in improving classification accuracy. In this paper we propose to use different objective functions, which are closer to the classification accuracy criterion, to replace the log-likelihood objective used in the MaxEnt model estimation process. Specifically, we propose a *Summation-Log K-norm* objective and a *Summation K-norm* objective. Our experiments conducted on two large volume topic classification dataset prove the effectiveness of our new objectives in improving topic classification performance on top of the state-of-art MaxEnt model.

Index Terms: topic classification, maximum entropy, k-norm.

1. INTRODUCTION

Topic classification refers to the technique of selecting a topic from a set of pre-defined candidates given a sentence or utterance. It has been widely used in the context of call routing in call center automation systems. Among the techniques that have been proposed for the topic classification, Maximum Entropy (MaxEnt) model [1][2][3] has been proven to be very effective. In many testing cases, MaxEnt Model has been shown to deliver the state-of-art performance in topic classification.

As proven by [1], the solution to the MaxEnt model is indeed an Exponential distribution which maximizes the likelihood of the training data, where the likelihood term contains the conditional (or posterior) probability of the topic given the sentence. Although this maximum likelihood formulation bears validity as it best accounts for the training data, and it produces better classification performance than many other topic classification methods, it's indeed not the exact function or criterion which will be used to judge the performance a topic classifier. Consequently, the MaxEnt model parameters which maximize the likelihood of the training data is sub-optimal in a sense that they have been developed under one objective while they will be evaluated based on another objective. To solve this "discrepancy" in the objectives, we propose some new objectives that could be used to estimate the model parameters of a exponential family distribution. These new objectives are closer to (some of them are indeed very close to) the real criterion of classification accuracy, and hence they are able to produce significant improvement in classification performance on top of the state-of-art MaxEnt model.

Our paper has been organized as follows: Section 2 describes the conventional MaxEnt method under the general framework of exponential family distribution. Section 3 describes our proposed new objectives that will be used as the optimization target of the exponential distributions. Our experimental results are presented in section 4, and we conclude the paper with discussions in section 5.

2. MAXIMUM ENTROPY MODEL: AN EXPONENTIAL DISTRIBUTION WITH MAXIMUM LOG-LIKELIHOOD ON THE TRAINING DATA

As described in [1][2], unlike the conventional Maximum Likelihood (ML) criterion, which is used to estimate a likelihood distribution of the data given the model $P(x|y)$, Maximum Entropy (MaxEnt) model produces a distribution of the conditional probability $P(y|x)$, if x is interpreted as the data sample (e.g. sentence as in topic classification), and y the class label (or topic) of the corresponding sentence. MaxEnt principle produces a statistical distribution of $P(y|x)$ to simultaneously satisfies the following two constraints:

1) This distribution should be "most uniform".

2) This distribution should in accord with the distribution or pattern of the important features presented in the training data.

The "uniform" criterion is defined through the measure of the conditional entropy $H(Y|X)$ of the probability distribution $P(y|x)$ as:

$$H(Y|X) = -\sum_{x,\,y} P(x)'P(y|x)LogP(y|x) \quad (1)$$

where $P(x)'$ is the empirical distribution of the training data x. The criterion of "most uniform" distribution can then be expressed as:

$$P(Y|X) = \underset{p}{argmax}\ H(Y|X) \quad (2)$$

The second criterion of distribution accordance is enforced by

constraining the expected value of the feature $f_i(x, y)$ calculated on the distribution $P(y|x)$ to be the same as what is observed in the training data:

$$\sum_{x,y} P(x,y)' f_i(x,y) = \sum_{x,y} P(x)' P(y|x) f_i(x,y) \quad (3)$$

where $P(x,y)'$ is the empirical joint distribution of the training data and topic label.

Incorporating the criterion of "most uniform" with the constraints of "same expectation" and probability distribution using Lagrangian, MaxEnt model then can be expressed as the following equation:

$$P(Y|X) = \underset{p}{argmax} \left(-\sum_{x,y} P(x)' P(y|x) Log P(y|x) + \sum_i \lambda_i \left(\sum_{x,y} P(x,y)' f_i(x,y) - \sum_{x,y} P(x)' P(y|x) f_i(x,y) \right) + \Upsilon \left(\sum_y P(y|x) - 1 \right) \right] \quad (4)$$

As shown in [1][2], the solution to Eq. (4) is indeed an exponential family distribution which maximizes the log-likelihood of the training data as:

$$P(Y|X) = \underset{p}{argmax} \sum_{x,y} P(x,y)' Log P(y|x) \quad (5)$$

where $P(y|x)$ is the exponential conditional probability distribution as:

$$P(y|x) = \frac{\exp\left(\sum_i \lambda_i f_i(x,y)\right)}{\sum_{y_j} \exp\left(\sum_i \lambda_i f_i(x,y_j)\right)} \quad (6)$$

From Eq. (5) and Eq. (6), it's clear that MaxEnt model estimation process is indeed an unconstrained optimization process with λ_i as the variable and O as the objective, as in Eq. (7):

$$\tilde{\lambda}_i = \underset{\lambda_i}{argmax} \; O = \underset{\lambda_i}{argmax} \sum_{x,y} P(x,y)' Log \left(\frac{\exp\left(\sum_i \lambda_i f_i(x,y)\right)}{\sum_{y_j} \exp\left(\sum_i \lambda_i f_i(x,y_j)\right)} \right) \quad (7)$$

3. REPLACING LOG-LIKELIHOOD OBJECTIVE WITH K-NORM BASED OBJECTIVES

It's clear from Eq. (7) that MaxEnt model estimation process is an optimization process, with O the log-likelihood of the training data as objective. Although this objective of training data log-likelihood has validity, it's not the exact criterion that will be used to evaluate the performance of a topic classifier. If we could replace this "log-likelihood" objective with some other objectives that are "closer" to the evaluation criterion, which is simply the classification accuracy or error rate, we should get better classification performance on top of the state-of-art performance of the MaxEnt model.

3.1 Classification Accuracy: The Ultimate Objective

Although we are using Bayes rule to make classification decision in topic classification, the final criterion for performance evaluation is the classification accuracy. If for each sentence x the label of the true topic and the topic with maximum classification score are y_c and $y(x)$ respectively, the classification accuracy C is simply:

$$C = \frac{\sum_x^N \delta(y_c, y(x))}{N} \quad (8)$$

where $\delta(x,y)$ is a unit function, it returns 1 when x and y matches, and 0 otherwise.

3.2 Shifting Conditional Probability Toward Classification Accuracy

The exponential conditional probability distribution of Eq. (6) can be treated as the posterior probability of topic y assuming all topics have flat prior probabilities. The problem with log-likelihood objective in Eq. (7) is that those correctly classified sentences may continue to increase their posterior probabilities during the optimization process without producing an increased classification accuracy. To solve this problem, we propose to add a "K-norm" factor on exponential component in both the numerator and denominator of the conditional probability expression in Eq. (6) as in Eq. (9):

$$P_k(y|x) = \frac{\exp\left(\sum_i \lambda_i f_i(x,y)\right)^K}{\sum_{y_j} \exp\left(\sum_i \lambda_i f_i(x,y_j)\right)^K} \quad (9)$$

When K equals 1, $P_k(y|x)$ is still the posterior probability of topic y (with flat prior probability assumption). But as we increase the value of K (e.g. from 1 to 100, 200 or even 1000), it gradually becomes the unit function $\delta(x,y)$ as described in Eq. (8). $P_k(y|x)$ will approximate to 1 when the term in numerator has the maximum value among all the potential topics y_j, which corresponds to a correct classification case if we take y as true topic label of the sentence x; and it will go to 0 when some other exponential terms instead of the numerator term has the maximum value, which is equivalent to a misclassification. Using this *"K-norm"* based posterior probability with larger K value is a better approximation to the classification accuracy than the posterior probability itself.

Another way to achieve an even better approximation to classification accuracy is to use *Summation* instead of *Summation-Log* as in Eq. (7) to accumulate individual probability terms. Since classification accuracy measure in Eq. (8) accumulates the performance of individual sentence through a *Summation* function, we hope this *Summation* of our *K-norm* objectives will achieve an even better approximation to the classification accuracy measure.

Based on the above discussions, we proposed to replace the log-likelihood objective used in the MaxEnt model of Eq. (7) with the following two objectives:

Summation-Log of *K-norm* as in Eq. (10)

$$\tilde{\lambda}_i = \underset{\lambda_i}{argmax}\, O = \underset{\lambda_i}{argmax} \sum_{x,y} P(x,y)' Log\left(\frac{\exp\left(\sum_i \lambda_i f_i(x,y)\right)^K}{\sum_{y_j} \exp\left(\sum_i \lambda_i f_i(x,y_j)\right)^K} \right) \quad (10)$$

, and *Summation* of *K-norm* as in Eq. (11)

$$\tilde{\lambda}_i = \underset{\lambda_i}{argmax}\; O = \underset{\lambda_i}{argmax} \sum_{x,y} P(x,y)' \frac{\exp\left(\sum_i \lambda_i f_i(x,y)\right)^K}{\sum_{y_j} \exp\left(\sum_i \lambda_i f_i(x,y_j)\right)^K} \quad (11)$$

3.3 Model Parameter Estimation for Our New Objectives

To estimate the model parameters λ_i under our new objectives, we have to first compute the derivative of the objective with respect to model parameters. The derivative of *Summation-Log* of *K-norm* and *Summation of K-norm* can be derived as the expression in Eq. (12) and Eq. (13) respectively:

$$\frac{\partial}{\partial \lambda_i} \sum_{x,y} P(x,y)' Log\left(\frac{\exp\left(\sum_i \lambda_i f_i(x,y)\right)^K}{\sum_{y_j} \exp\left(\sum_i \lambda_i f_i(x,y_j)\right)^K} \right)$$
$$= \sum_{x,y} P(x,y)' \cdot K \cdot \left\{ f_i(x,y) - \frac{\sum_{y_j} \left[\exp\left(\sum_i \lambda_i f_i(x,y_j)\right)^K \cdot f_i(x,y_j) \right]}{\sum_{y_j} \exp\left(\sum_i \lambda_i f_i(x,y_j)\right)^K} \right\} \quad (12)$$

Based on the derivative expression of our new objective with respect to the model parameter, we can carry optimization process (e.g. BFGS [4]) to find the optimal solution which maximizes our objective and hence achieves better topic classification accuracy.

4. EXPERIMENTAL RESULTS

In order to test the effectiveness of our new objectives of *Summation-Log K-norm* and *Summation K-norm*. We carried out some experiments using two different topic classification data sets. One is the conversational data collected from a tele-communication domain, which we denote it as "TC", that contains about 200000 sentences in the training set and about 20000 sentences in the testing set. We have allocated around 20000 sentences from the training set as the development set, which will be used to control the termination of iterative updating procedure. The total number of topics in the "TC" dataset is 482, and the number of words in the vocabulary is a little bit over 7000. The second dataset came from a courier NLU application, which we named it as "CV". It contains about 65000 sentences in the training set and 7000 sentences in the testing set. We also split the training data into 60000 and 10000 as training and development set. The vocabulary size of the second dataset is around 4000, and number of topics is 31.

Error rate (%)	*MaxEnt*	*Summation-Log*	*Summation*
TC dataset	9.88	8.69	8.92
CV dataset	10.2	9.14	9.46

Table 1. Topic classification error rate comparison between MaxEnt model and our new objectives. Value of *K* in the *K-norm* is 100.

To estimate the model parameters under the new objective, we modified the L-BFGS optimization code as in [4]. We select the model which produces the best classification result on the validation set for testing.

4.1 Comparison of Classification Results between MaxEnt Model and Our New Objectives

The first performance comparison was between MaxEnt model and exponential distribution model estimated using our new objectives, as been illustrated in Table 1. The specific value of *K* in *K-norm* is 100. As can be seen from Table 1, we achieved 12% relative improvement in TC dataset, and more than 10% relative improvement in CV dataset, which is quite significant, especially considering the fact that this improvement was achieved on top of the state-of-art MaxEnt model.

4.2 Comparison between Classification Results of Using Different K-norm Values

Another experiment we carried was to compare the topic classification performance with different *K* values in the *K-norm*. We tested the specific value of 100, 200, and 1000 of *K-norm* on both *Summation-Log* and *Summation* objectives, and their results are illustrated in Table 2. Based on the experimental results, whether we use 100, 200 or 1000 doesn't make too much difference on the classification performance. We believe this is a reasonable observation. According to Eq. (12) and Eq. (13), the specific value of *K* affects both the "conditional" probability term and the scale of the model parameter derivative. When $K \gg 1$, which we believe is a valid condition for K=100, 200, and 1000, the sensitivity of the "conditional probability" term on *K* value is insignificant, and the only noticeable impact from *K* value on the model parameter derivative is the scale of the derivative, which doesn't affect the result of optimization process very much. For this reason, unless otherwise specified, all our experiments take *K* value as 100.

4.3 Comparison in Log-Likelihood Between MaxEnt Model and Our New Objectives

$$\frac{\partial}{\partial \lambda_i} \sum_{x,y} P(x,y)' \frac{\exp\left(\sum_i \lambda_i f_i(x,y)\right)^K}{\sum_{y_j} \exp\left(\sum_i \lambda_i f_i(x,y_j)\right)^K} = \sum_{x,y} P(x,y)' \frac{\exp\left(\sum_i \lambda_i f_i(x,y)\right)^K}{\sum_{y_j} \exp\left(\sum_i \lambda_i f_i(x,y_j)\right)^K} \cdot K \cdot \left\{ f_i(x,y) - \frac{\sum_{y_j} \left[\exp\left(\sum_i \lambda_i f_i(x,y_j)\right)^K \cdot f_i(x,y_j) \right]}{\sum_{y_j} \exp\left(\sum_i \lambda_i f_i(x,y_j)\right)^K} \right\} \quad (13)$$

Error rate on TC dataset (%)	*Summation-Log*	*Summation*
K=100	8.69	8.92
K=200	8.63	8.92
K=1000	8.72	8.94

Table 2. Topic classification error rate comparison between different *K-norm* values on TC dataset.

Our original motivation of using *K-norm* objectives instead of the posterior probability is that we believe the optimization effort may be wasted in the later case as the increase of objective function value may be introduced by those already correctly-classified sentences. *K-norm* based objectives should compensate this bias by balancing the optimization process toward those misclassified sentences.

To verify the validity of this motivation, we did another experiment and compare the log-likelihood of various objectives as the following: We first split the training set into two parts as those "correctly-classified" sentences and those "misclassified" sentences produced by an intermediate MaxEnt model, which has been generated in the middle of the Improved Iterative Scaling process. Then we continue carry optimization on the training data using both Log-likelihood objectives as in MaxEnt and our *K-norm* objectives. We compared the changes in both the log-likelihood and number of correctly or in-correctly classified sentences on both sentence sets, and the results are listed in Table 3.

Table 3 unveils some interesting observations. If we compare the changes on those originally "correctly-classified" sentences that have been introduced by the continuous optimization of MaxEnt model, we observed an increase in the log-likelihood (i.e. +1233.13), while at the same time a decrease in classification performance (i.e. 86 sentences which were originally classified correct are now become "misclassified"). We believe this confirms our original hypothesis that by using "log-likelihood" as optimization criterion, the optimization process may be biased toward some of the "correctly-classified" sentences by improving their likelihood while sacrifice those "on the edge" or "misclassified" sentences. On the other hand, using our *K-norm* based objectives may correct this bias, as they force the optimization process toward those "misclassified" or "on-the-edge" sentences, which is clearly reflected from Table 3 by comparing the changes in log-likelihood and classification error reductions in those originally "misclassified" sentences between MaxEnt model and our *Summation-Log 100-norm.*

Change of Log-likelihood	*Correctly Classified Sentences*	*Misclassified Sentences*
MaxEnt	+1233.13	+226.37
K-norm, Summation-Log(K=100)	-10309.3	+2042.88

Sentence Count Changes	*Correctly Classified Sentences*	*Misclassified Sentences*
MaxEnt	-86	-744
K-norm, Summation-Log(K=100)	-862	-4578

Table 3. Changes in the Log-likelihood and count of correctly and in-correctly classified sentences in the training data produced by an intermediate MaxEnt model on TC dataset.

Inspired by this observation, we did another experiment which was to force the MaxEnt model to only optimize those originally "misclassified" sentences on the TC set to see whether it improves the overall classification performance. This approach reduced the topic classification error rate into 9.45% (as opposed to 9.88% in the original MaxEnt model), but is still behind the 8.69% achieved by using our new K-norm objectives.

5. DISCUSSIONS AND CONCLUSIONS

In this paper we try to improve the topic classification accuracy of the state-of-art MaxEnt model by replacing its original log-likelihood objective with our new *K-norm* based objectives. Because the model estimation process of a MaxEnt model is indeed an unconstrained optimization process of an exponential family distribution, we can easily plug-in our new objectives and carry optimization with minimum additional effort. Our experimental results conducted on two large volume call routing dataset proved the effectiveness of our new objectives in improving classification performance on top of the MaxEnt model, as we achieved more than 12% and 10% relative reduction in classification error rate while the computation effort remains the same. Our experimental results regarding the relation between change of log-likelihood and reduction in classification error rate also confirm the validity of our new objectives.

One thing we did notice from experimental results was the fact that *Summation K-norm* objective actually performs worse than the *Summation-Log K-norm* objective. Although we are currently still investigating this, we suspect this could be introduced by the following fact: For the *Summation* accumulation function, the effect of changing a classification result on a single sentence to the overall objective is in the range of -1 to 1, which is a very small number compared to the absolute value of the objective (especially considering the large number of training sentences). This may in turn cause the optimization process to "ignore" some of those on the "edge" sentences, as the change of their classification status will not produce too much change to the overall value of the objective. *Summation-Log* on the other hand can somehow avoid this problem, as switching of classification result (e.g. from $Log1$ to $Log0$) on a single sentence may produce a big impact on the overall value of the objective.

6. REFERENCES

[1] A. L. Berger, S. A. Della Pietra, and V. J. Della Pietra, "A maximum entropy approach to natural language processing," Computational Linguistics, Vol. 22, No.1, pp. 39–72, March 1996.

[2] K. Nigam, J. Lafferty, and A. McCallum. "Using maximum entropy for text classification", In IJCAI-99 Workshop on Machine Learning for Information Filtering, pp. 61-67, 1999

[3] C. Chelba, M. Mahajan, A. Acero, "Speech Utterance Classification", Proceedings of ICASSP 2003, 2003.

[4] C. Zhu, R. H. Byrd and J. Nocedal. "L-BFGS-B: Algorithm 778: L-BFGS-B, FORTRAN routines for large scale bound constrained optimization", ACM Transactions on Mathematical Software, Vol 23, No. 4, pp. 550 - 560, 1997