Minimum Entropy Estimation of Hierarchical Random Graph Parameters for Character Recognition

Ho-Yon Kim
Electronics and Telecommunications Research Institute
hykim@etri.re.kr,
Jin-H. Kim
Department of Computer Science & Center for AI Research, KAIST
jkim@ai.kaist.ac.kr

Abstract

In this paper, we propose a new parameter estimation method called minimum entropy estimation (MEE), which tries to minimize the conditional entropy of the models given the input data. Since there is no assumption in MEE for the correctness of the parameter space of models, MEE will perform not less than the other estimation methods such as maximum likelihood estimation (MLE) and maximum mutual information estimation (MMIE), under the condition that the training data size is large enough. In the experiments, the three estimation methods are applied to the parameter estimation of hierarchical random graphs so that their estimation performance can be compared with each other.

1. Introduction

In this paper, we will review the parameter estimation problem in the information theoretic view, and present an estimation criterion called *minimum entropy estimation* to estimate stochastic model parameters.

The three estimation methods introduced and examined include *maximum likelihood estimation* (MLE), *maximum mutual information estimation* (MMIE), and *minimum entropy estimation* (MEE), which is the new estimation method derived in this paper.

The rest of this paper is organized as follows. In section 2, we will summarize information theory and the concepts of MLE and MMIE. Then, the new estimation method MEE will be introduced in section 3. Experimental results for the application to handwritten Hangul recognition are given in section 4, followed by conclusion in section 5.

2. Background

2.1 Entropy and mutual information

According to information theory [1], the uncertainty of a source X with the probability P(X=x) called the entropy of X is measured with the equation

$$H(X) = -\sum_{x} P(X=x)\log P(X=x), \tag{1}$$

and the conditional entropy of X given Y is defined as

$$H(X\,|\,Y) = -\sum_{x,y} P(X=x, Y=y)\log P(X=x\,|\,Y=y) \tag{2}$$

Then, the *average mutual information* between X and Y can be obtained by subtracting $H(X|Y)$ from $H(X)$:

$$\begin{aligned} I(X;Y) &= H(X) - H(X\,|\,Y) \\ &= \sum_{x,y} P(X=x, Y=y)\log\frac{P(X=x, Y=y)}{P(X=x)P(Y=y)} \end{aligned} \tag{3}$$

2.2 Parameter estimation of probabilistic models

The recognition in probabilistic models is to find a model $\hat{M}$ maximizing the conditional probability $P(\hat{M}\,|\,x)$ given an instance x, which is formulated as

$$\arg\max_{\hat{M}} P(\hat{M}\,|\,x) = \arg\max_{\hat{M}} \frac{P(x,\hat{M})}{P(x)}. \tag{4}$$

Since the true probability distribution of $P(x,\hat{M})$ is unknown, $P_{\Theta}(x,\hat{M})$, the probability distribution approximated by the parameter vector Θ, is used for the unknown probability distribution. In the view of information theory, to get much more information about M when $X=x$ is observed, we have to adjust the parameter vector Θ so that the uncertainty of M given X becomes as small as possible. In other words, to enhance recognition performance, we have to find Θ minimizing $H_{\Theta}(M|X)$, the conditional entropy of M given x, parameterized by Θ.

0-7695-0750-6/00 $10.00 © 2000 IEEE

The objective function, $H_\Theta(M|X)$, to be minimized is written as

$$H_\Theta(M \mid X) = H_\Theta(M) - I_\Theta(M;X), \tag{5}$$

in which

$$H_\Theta(M) = -\sum_m P(M=m)\log P_\Theta(M=m), \tag{6}$$

$$I_\Theta(M;X) \tag{7}$$
$$= \sum_{m,x} P(M=m, X=x)\log\frac{P_\Theta(M=m, X=x)}{P_\Theta(M=m)P_\Theta(X=x)}.$$

Some objective functions to estimate probabilistic model parameters can be induced from the equation $H_\Theta(M|X)$. Especially, we are concerned with the two representative estimation methods: *maximum likelihood estimation* (MLE) and *maximum mutual information estimation* (MMIE). In the following section, we will review the meaning of MLE and MMIE with respect to information theory, and then propose the other estimation method. Simply speaking, the proposed estimation method, called *minimum entropy estimation* (MEE), differs from MMIE in that $H_\Theta(M)$, the entropy of models, is also minimized.

2.3 Maximum likelihood estimation

Maximum likelihood estimation is to choose a parameter vector $\hat{\theta}_i$ maximizing the outcome probability of X given model M_i:

$$\hat{P}_{\theta_i} = \max_{\theta_i} P_{\theta_i}(X \mid M_i). \tag{8}$$

From the equation (8) we can notice that MLE finds the model parameters maximizing the outcome probabilities of the samples given a model. That is, each model is separately estimated as close to the true parameters as possible. The criterion of MLE can be derived from the objective function $H_\Theta(M|X)$ on the two assumptions that 1) the entropy of M is not related to the parameter vector Θ, and 2) the family of distributions of X given M is known. On the first assumption, $H_\Theta(M)$ in the equation (5) can be ignored, and on the second assumption, $H_\Theta(X)$ in $I_\Theta(M;X)$ can be ignored, which results in leaving the criterion of MLE.

2.4 Maximum mutual information estimation

The objective function of *maximum mutual information estimation* (MMIE) [4] can be derived from the equation (5) on the assumption that $H_\Theta(M)$, the entropy of the model M, is independent on the parameter vector Θ. This assumption is often replaced with the assumption that the prior probability of M is equal to the true probability, or can be computed with a large database. The optimization criterion of MMIE is

$$\hat{I}_\Theta = \max_\Theta \log\frac{P_\Theta(M=m, X=x)}{P_\Theta(M=m)P_\Theta(X=x)} \tag{9}$$

$$= \max_\Theta\{\log P_\Theta(X=x \mid M=m) - \log P_\Theta(X=x)\}$$

$$= \max_\Theta\{\log P_\Theta(X=x \mid M=m)\} - \log\{\sum_m \{P_\Theta(X=x \mid M=m)P(M=m)\}\}$$

As shown in the equation (9), the parameter vector Θ is chosen to maximize $I_\Theta(M;X)$, the *average mutual information* between a set of models and the training samples. In other words, in MMIE, the models are trained at the same time so as to maximize the discrimination power of each model. In this sense, the major difference of MMIE from MLE is that MMIE tries to not only enlarge the outcome probabilities of the samples given the target model, but also reduce the outcome probabilities of the samples given the others.

3. Minimum entropy estimation

Although MMIE doesn't assume all the premises MLE has, it still has an assumption that $H_\Theta(M)$, the entropy of the model M, is independent on the parameter vector Θ. Since, in the previous studies, $P_\Theta(M=m)$ has been regarded as the prior probability of model m and assumed to be known or easily computed from a large database, it has seemed to have no relation to the parameter vector Θ. This is not always true.

Let's define as $P_I(M=m)$ the expected prior probability of model m that is estimated by means of MMIE. Then it can be represented with the summation of marginal probabilities such that

$$P_I(M=m) = \sum_x P_I(M=m, X=x)$$

$$= \sum_x P_\Theta(X=x \mid M=m)P(X=x) \tag{10}$$

From equation (10), it is certain that $P_I(M=m)$ varies according to the parameter vector Θ, and the prior probabilities acquired from the estimated models are not equal to the real prior probabilities if the assumed probability distribution is incorrect. Thus, estimating the prior probability distribution from training data, we can compensate the biased prior probabilities.

In this respect, we propose a new estimation method called *minimum entropy estimation* (MEE), which tries to minimize the conditional entropy of M given X. Since the criterion of MEE is to minimize directly $H_\Theta(M|X)$, MEE is the most general and has no assumption that may be incorrect. The optimization criterion of MEE is as follows.

$$\hat{E}_\Theta = \min_\Theta\left\{-\log P_\Theta(M=m) - \log\frac{P_\Theta(M=m,X=x)}{P_\Theta(M=m)P_\Theta(X=x)}\right\} \tag{11}$$

$$= \max_\Theta\left\{\log P_\Theta(M=m) + \log\frac{P_\Theta(M=m,X=x)}{P_\Theta(M=m)P_\Theta(X=x)}\right\}$$

$$= \max_\Theta\{\log P_\Theta(M=m)\} + \max_\Theta\{\log P_\Theta(X=x\,|\,M=m) - \log P_\Theta(X=x)\}$$

$$\hat{E}_\Theta = \max_\Theta\{\log P_\Theta(M=m)\} + \hat{I}_\Theta \tag{12}$$

As shown in the equation (12), MEE tries to maximize P_I $(M=m)$ in addition to what $\hat{I}_\Theta$ maximize. We can interpret the meaning of the estimated $P_\Theta(M=m)$ as follows. Let m be a model, and p be a true prior probability of $P(M=m)$. Then, if the expected prior probability P_I $(M=m)$ is larger than the true prior probability $P(M=m)$, $P_\Theta(M=m)$ gets smaller, and vice versa. In this way, MEE tries to minimize the entropy of M given X, which gives maximum information for the given X.

Another advantage of MEE is that it enables us to consider different features for different models. For example, we may use discrete probability distribution for a model, while continuous probability distribution for another model. We may also use different number of probability distributions for each model. This is possible because $P_\Theta(M=m)$ plays a role in adjusting differences of the scales of the probability distributions.

4. Experimental results

In this section, the experiments to test the three estimation methods are described. The stochastic model used in the experiments is the hierarchical random graph [7] that had been introduced to model handwritten Hangul characters. The experiments for the parameter estimation were conducted with handwritten Hangul characters in the KU-1 database [3] containing 1000 sets of 520 unconstrained handwritten characters with different labels. Among the database, we used 200 sets for training, and 100 sets from the rest for testing, which are enough to estimate the proposed model as will be validated in section 4.2.

4.1 Baseline model

Since MLE is much faster than MMIE and MEE, the baseline models used in MMIE and MEE is obtained from MLE-trained models, where initial parameters are assigned manually. Series of experiments have been carried out to get the better-estimated parameters of the baseline models. At first, we estimated the parameters with a small set of training samples by applying both MLE and *parameter smoothing*. Then, with a large set of training samples, we obtained better-estimated parameters.

4.2 Determination of the training data size

According to Brown [5] and Niles [6], the performance of MLE and MMIE depends on the amount of the training data. They say that MLE is better than MMIE when the model fits the data well and training data is limited, while MMIE is better when the model has a poor fit. Thus, to analyze the performance of the estimation methods more precisely, it is necessary to measure the amount of the training data enough to estimate the model parameters.

To ascertain how much training data is needed to estimate the parameters of the proposed Hangul model with a lot of discrete probability distributions to be estimated, we investigated the change of recognition rate as a function of the size of the training data. In general, when the training data is insufficient to estimate parameters, recognition rate of test data becomes low, but that of the training data becomes high. This situation is called specialization. On the other hands, as the training data keeps going on increasing, assuming that the data are independently and identically distributed, the recognition rate of test data becomes higher bounded by that of the training data while the recognition rate of the training data becomes lower bounded by that of test data. Ideally, as the size of training data goes infinity, the two recognition rates will converge on a same value.

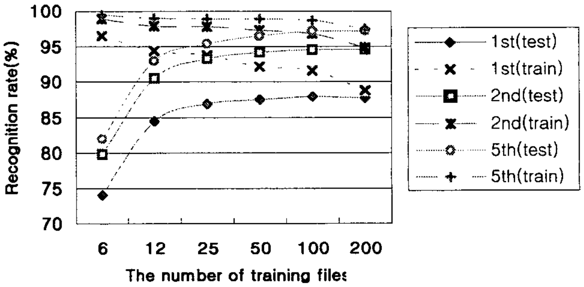

Figure 1: Recognition rate for training data and test data as a function of the number of the training data files

In our experiment, the recognition rate of test data is observed as the training data is increased roughly two times. 20 sets from the SERI database are used for testing and 6, 12, 25, 50, 100, and 200 sets are used for training in the order of the size. So, 10400 characters are used for preliminary testing, and from 3120 to 104000 characters are used for training. Experimental results are given in Figure 1, where recognition rates for the training data and test data are shown with their recognition rates up to the second and the fifth candidates. As shown in Figure 1, the recognition rate of the training data gets lower while that of test data gets higher as the training data size becomes larger. With respect to the convergence of the recognition rate for test data and the small difference of the recognition rate between the training ones and test ones, 200 sets from the database seems to be enough to estimate the proposed Hangul model parameters. Therefore, we used 200 sets for training, and 100 sets from the rest for testing. That is, 104000 characters were used for parameter estimation, and 52000 characters were used for testing.

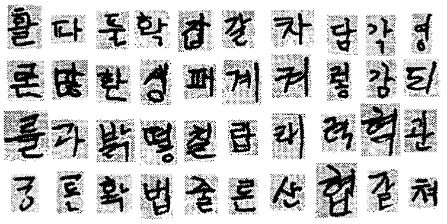

Figure 2: Examples of correctly recognized data

4.3 Comparison of MLE, MMIE and MEE

To compare the performance of MLE, MMIE, and MEE, 200 sets of the SERI database are used for training and 100 sets are used for testing. When the number of recognition targets is 2350, the recognition rates of each method are compared in Table 1. It is clear that MMIE outperforms MLE. Since the training data is almost enough to estimate the model parameters, it can be expected that MMIE is better than MLE.

Table 1: Comparison of MLE, MMIE, and MEE

Estimation method	training data : 104000 char. 1st (2nd, 5th)	test data : 52000 char. 1st (2nd, 5th)
MLE	84.09 (91.86, 95.97)	82.73 (91.29, 95.7)
MMIE	87.23 (93.49, 96.50)	86.10 (92.99, 96.25)
MEE	87.57 (93.69, 96.58)	86.31 (93.15, 96.36)

On the other hands, the performance of MEE is not much better than that of MMIE. In our experimental context, the difference of the prior probabilities of each model may not be large. Furthermore, since we applied the prior probability not to character models but to sub-character models [7], the effect of the prior probability may be small. However, at any rate, we can say that MEE can compensate the incorrect prior probability stemming from incorrect model distribution.

5. Conclusion

In this paper, we reviewed parameter estimation methods of statistical models in an information theoretic view, and applied them to the parameter estimation of a hierarchical random graph. The two representative estimation methods including maximum likelihood estimation and maximum mutual information estimation, which have been used for parameter estimation of HMM, were implemented and examined.

In addition, a new estimation method called *minimum entropy estimation* (MEE) has been proposed, having no assumption of the model correctness. The optimization criterion of MEE is very similar to that of MMIE except that MEE also tries to maximize the prior probability of models, as well as mutual information between training samples and models. Since MEE is derived from entropy theory without any premises, it will perform as well as any other estimation methods if we have enough training data. The experiments to compare the three estimation methods have shown that MEE is the best, and MMIE outperforms MLE as Brown [5] pointed out, when training data is large enough.

By applying the estimation criteria to the parameter estimation of the hierarchical random graph for stochastic modeling of handwritten characters, we improved the recognition performance of a handwritten Hangul recognition system.

References

1. C. E. Shannon "A mathematical theory of communication," Bell System Technical Journal, 27 (1948)
2. A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Royal Statistical Society* 39, 1-38 (1977).
3. D.-I. Kim and S.-W. Lee, "An automatic evaluation of handwriting qualities for off-line handwritten Hangul character database KU-1", *Proc. of the 25th Korea Information Science Society Conference* 25(1), 707-709 (1998).
4. L. R. Bahl, P. F. Brown, P. V. de Souza and R. L. Mercer, "Maximum mutual information estimation of hidden Markov model parameters for speech recognition," *Proc. ICASSP-86, Tokyo*, 49-52, (Apr. 1986)
5. P. F. Brown, "The acoustic-modeling problem in automatic speech recognition", CMU-CS-87-125.
6. L. T. Niles, H. F. Silverman and M. A. Bush, "Neural networks, maximum mutual information training, and maximum likelihood training," Proc. ICASSP-90, (Apr. 1990)
7. H. Y. Kim and J. H. Kim, "Handwritten Korean character recognition based on hierarchical random graph modeling," Sixth International Workshop on Frontiers in Handwriting Recognition, 577-586, (1998)