Estimate The Source Structure Through Communication

Shiro IKEDA and Kaoru NAKANO
Department of Mathematical Engineering and Information Physics
University of Tokyo
7-3-1, Hongo, Bunkyo-ku, Tokyo, 113, Japan
E-mail: shiro@bcl.t.u-tokyo.ac.jp

ABSTRACT

In order to categorize a set of data which consists of some categories, it is important to know the probability distribution of the data of each category. Using these probability distributions, we can classify the data. In most cases, such as speech and image recognition problems, the data for training are categorized in advance. But there are some cases that only the uncategorized data are available. For example, when a baby learns phonation, it is hard to give it the samples of each phone separately. It learns the number of the phones and how to phonate each of them through observing only the uncategorized data and making communication with the parents or the teacher. In this article, the authors give an algorithm for this situation. In the algorithm, the distribution of whole data is described with a finite mixture model. The model can observe only the uncategorized data but can make a kind of communication with the teacher (which is the probability source). The parameters of the model are estimated using the EM algorithm and the number of the categories are determined through a communication with the teacher. A numerical simulation of a simple image recognition problem is given.

1. Introduction

Suppose that there is a set of data which consists of some categories, and that we want to know the categories of new data. For this purpose, it is important to know the probability distribution of the data of each category. Using the probability density functions, their categories can be judged. We can find this kind of problems in constructing speech or image recognition systems. In such cases, we usually have some data whose categories are known in advance. Therefore, each probability distribution can be estimated respectively.

But there are different cases that the categorized data are not available. In such cases, one good approach is to make a model for whole data with finite mixture model, and estimate its parameters[3].

Here is an example. When a baby learns the phonation, it doesn't know how many phones there are, nor how to phonate each of them. The parents and the teacher cannot tell it the number of the phones and it is hard to give it the samples of each phone separately. This corresponds to what is mentioned above. What is different is that the baby can make a kind of communication with the teacher. It phonates ambiguous sounds according to its own model, and the teacher corrects them.

In this article, the authors treat a problem that only uncategorized data are available, but the model, which is described with a finite mixture model, can make a kind of communication with the teacher which is a probability source (we call this the source). If the number of the categories is assumed, then the parameters of the model can be estimated using the EM algorithm described in Section 2. The number of the categories are determined through communication with the source. The concept of the algorithm is shown in Section 3. The algorithm is applied for a simple image recognition problem, and made a good result. The results are shown in Section 4.

2. EM algorithm

The data $\{x\}$ is supposed to have some categories, but we cannot observe them. In such situation, one good approach to estimate probability distribution of each categories, is to describe the distribution of whole data with a finite mixture model, and estimate the parameters. Then each component distribution will correspond to each category's distribution. Because the number of the categories is unknown, we assume the number is m, each category is denoted by z, $(z = 1, \cdots, m)$. The component density function of category z is $p_z(x|\theta_z)$ and the mixing weights are π_z, $(\sum_z \pi_z = 1)$. Then the distribution of x is described as finite mixture model [3]

$$p(x|\theta) = \sum_{z=1}^{n} \pi_m p_z(x|\theta_z), \tag{1}$$

where $\theta = (\pi_1, \cdots, \pi_m, \theta_1, \cdots, \theta_m)$. When we have a new datum x, we should decide that the category it belongs to is z which maximizes $\pi_z p_z(x|\theta_z)$.

When we have $\{x_1, x_2, \cdots, x_N\}$ for estimating θ, we can use maximum likelihood method where θ is chosen to maximize $\prod_{s=1}^{N} p(x_s|\theta)$ which is equivalent to maximizing $\sum_s \log p(x_s|\theta) = \sum_s l(x_s|\theta)$. But in this case, we have the hidden variable z. Therefore, it is hard to find the θ directly by solving it. We can use the EM algorithm to solve this problem [2].

The EM algorithm generates, from some initial point θ^0, a sequence $\{\theta^t\}$ of estimates. Each iteration consists of the following double step:

- E-step
  Evaluate $E_{\theta^t}[l(x, z|\theta)]_{x_s}$, which is

$$Q(\theta, \theta^t) = \frac{1}{N} \sum_{s=1}^{N} \left\{ \sum_z l(x_s, z|\theta) p(z|x_s, \theta^t) \right\} \tag{2}$$

- M-step
  Find the θ^{t+1} which maximizes $Q(\theta, \theta^t)$,

$$\theta^{t+1} = \arg\max_{\theta} Q(\theta, \theta^t). \tag{3}$$

where

$$p(z|x, \theta) = \frac{p(z, x|\theta)}{p(x|\theta)} = \frac{\pi_z p_z(x|\theta_z)}{\sum_{z'} \pi_{z'} p_{z'}(x|\theta_{z'})}$$

$$l(x, z|\theta) = \log p(x, z|\theta) = \log \pi_z + \log p_z(x|\theta_z).$$

It can easily be proved that $\sum_s l(x_s|\theta^{t+1}) \geq \sum_s l(x_s|\theta^t)$[2]. Thus, we can obtain the maximum likelihood estimators by iterating these two steps.

3. Determine the number of the categories

In section 2, it is shown that if the number of the categories is defined and the structure of the component densities are given, we can estimate the parameters by the EM algorithm. We presume that the structure of each component densities are given. Then, the remaining problem is to determine the number of the categories.

To estimate the number of the categories, we can use some information criteria, such as AIC[1]. But it is hard to believe that a baby is calculating AIC when it learns phonation. There must be some kind of criteria, but it might not exactly the same as AIC.

The authors assume that the number of the categories are determined by communicating with the source. Starting from the mixture model which has only 1 category, the model makes communication, and decides whether to make more categories. Iterating this, the model determines the number of the categories.

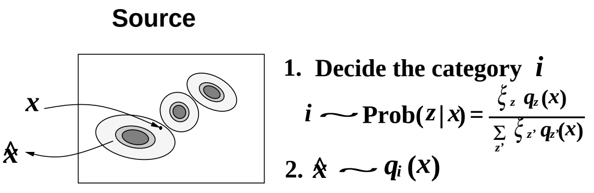


Fig. 1. Behavior of The Probability Source

First, we describe the behavior of the source and the way the model makes communication with the source. As mentioned above, the source doesn't give the information of the categories directly but give a new datum $\hat{x}$ according to his own distribution $q(x|\psi) = \sum_{i'} \xi_{i'} q_{i'}(x|\psi_{i'})$ when a datum x is given. The source's behavior is shown in Figure 1 and is described as follows:

1. Receives x from the model.
2. Judge the category i to which the datum is belonging based on its own probability distribution $q(x|\psi) = \sum_{i'} \xi_{i'} q_{i'}(x|\psi_{i'})$.
3. Generate $\hat{x}$ according to probability density function $q_i(x|\psi_i)$, and give it to the model.

The model makes communication with this source as follows:

1. Choose the category z to ask the source at random, according to the mixing weights π_z. Then generate x according to the density function $p_z(x|\theta_z)$, and ask it to the source.
2. Receive the $\hat{x}$ from the source.

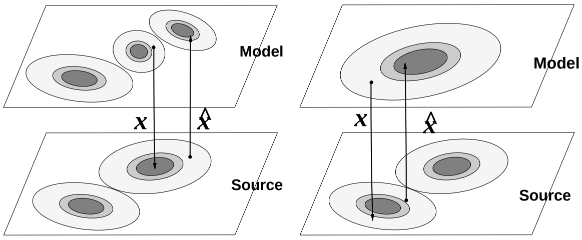


Fig. 2. Behavior of $d(x, \hat{x})$

We are now going to show that by observing x and $\hat{x}$), we can determine the number of the categories.

By Maximum Likelihood Estimate, we try to maximize $\sum_s l(x_s|\theta^{t+1})$. This means we are trying to maximize, $E[l(x|\theta)] = \int q(x) \log p(x|\theta) dx$. This can also mean we are trying to minimize,

$$\begin{aligned} & D(q(x), p(x|\theta)) \\ = & \int q(x) \log q(x) dx - \int q(x) \log p(x|\theta) dx \\ = & \int q(x) \log \frac{q(x)}{p(x|\theta)} dx. \end{aligned} \tag{4}$$

Here $D(,)$ is the Kullback-Leibler Divergence.

But now, we have the data $\{x\}$ and $\{\hat{x}\}$. Therefore we should consider the marginal distribution of $\{x\}$ and $\{\hat{x}\}$. Thus we should choose the model which would minimize

$$\begin{aligned} D_{\hat{x}x}(q,p) &= D(q(x,\hat{x}), p(x,\hat{x}|\theta)) \\ &= \int q(x,\hat{x}) \log \frac{q(x,\hat{x})}{p(x,\hat{x}|\theta)} dx d\hat{x}. \quad (5) \\ &= \int q(x,\hat{x}) \log q(x,\hat{x}) dx d\hat{x} \\ &\quad - \int q(x,\hat{x}) \log p(x,\hat{x}|\theta) dx d\hat{x} \quad (6) \end{aligned}$$

To make our idea clear, it is better to describe the behavior of $D_{\hat{x}x}(q,p)$. $D_{\hat{x}x}(q,p)$ is a kind of distance between probabilistic distributions. Therefore, if the two distributions are “close”, $D_{\hat{x}x}(q,p)$ will be small, otherwise, it is large. Suppose that the source's probability distribution is also a finite mixture model. If the categories of the model is less than the source, the source can represent the data more strictly than the model. On the other hand, if the model has more categories than the source, the model seems to have some categories which do not exist in the source. In both cases the value $D_{\hat{x}x}(q,p)$ will be larger than the model which has correct number of categories(Figure 2). Therefore, by observing the value of $D_{\hat{x}x}(q,p)$, we can get a kind of criteria for selecting the number of the categories.

But here is the problem. We cannot know the value of $D_{\hat{x}x}(q,p)$. The first term of (6) is common to every model and we do not have to calculate. The second term can be written as

$$\begin{aligned} &\int q(x,\hat{x}) \log p(x,\hat{x}|\theta) dx d\hat{x} \\ &= \int q(\hat{x}|x) q(x) \log p(x,\hat{x}|\theta) dx d\hat{x}. \quad (7) \end{aligned}$$

By substituting q(x) of (7) with $p(x|\theta)$, we can estimate the value of (7).

Consequently, we get the algorithm to estimate the number of categories.

1. Define the number of the categories and estimate the parameters by EM algorithm.
2. With the parameter θ^*, ask the source $\{x\}$ and receive $\{\hat{x}\}$
3. Estimate the value

   $$\int q(\hat{x}|x) p(x) \log p(x,\hat{x}|\theta) dx d\hat{x}$$

   as $\sum_s \log p(x_s, \hat{x}_s|\theta)$. And estimate the number of the categories.

4. Simulation

Figure 3 schematically shows the problem of the simulation. Each pattern is an n bits pattern of $\{1,0\}$. The source has only m patterns $\{M_z = (M_z^1, \cdots, M_z^n)^t : z = 1, \cdots, m\}$ of 2^n. These m patterns are corresponding to the categories. For each i, there defined an error rate e_z ($e_z < 0.5$) with which each bit turns around independently. Thus, we cannot observe the categories directly. Through communication with the source, the correct patterns, mixing weights and error rates are estimated(Figure 3). The conditional probability densities of a pattern $x = (x^1, \cdots, x^n)^t$ to category z is

$$p_z(x|\theta_z) = {e_z}^{d(x,M_z)} (1-e_z)^{(n-d(x,M_z))}. \quad (8)$$

where, $d(x, M_z)$ is

$$d(x, M_z) = \sum_{i=1}^{n} (x^i - M_z^i)^2. \quad (9)$$

and the probability distribution $p(x|\theta)$ is

$$p(x|\theta) = \sum_z \pi_z p_z(x|\theta_z). \quad (10)$$

Now we have to know the form of the distribution $p(\hat{x}, x|\theta)$.

$$\begin{aligned} & p(x,\hat{x}|\theta) = p_\theta(\hat{x}|x) p(x|\theta) \quad (11) \\ &= \sum_z p_\theta(\hat{x}, z|x) p(x|\theta) \\ &= \sum_z p_\theta(\hat{x}|z) p_\theta(z|x,\theta) p(x|\theta) \\ &= \sum_z p_z(\hat{x}|\theta_z) \frac{p_\theta(x,z|\theta)}{p(x|\theta)} p(x|\theta) \\ &= \sum_z \pi_z p_z(\hat{x}|\theta_z) p_z(x|\theta_z) \\ &= \sum_z \Big\{ \pi_z {e_z}^{(d(x,M_z)+d(\hat{x},M_z))} \\ &\quad \times (1-e_z)^{(2n-d(\hat{x},M_z)-d(\hat{x},M_z))} \Big\} \quad (12) \end{aligned}$$

In the simulation, the source has the categories shown in Figure 4. Each pattern is 9×9 ($n = 81$) bits pattern. For each category, as shown in the figure, the mixing weights and the error rates are different. We prepare 100 data for parameter estimation in advance. For parameter estimation, we used the EM algorithm. We do not give the concrete expression of the algorithm, but it can easily be derived from (2) and (3).

If the number of the categories is assumed to be 3, the model is estimated as shown in Figure 5. In this case, it is shown that the model tries to cover more than one category of the source's with one category by letting error rates and mixing weights large.

On the other hand, Figure 6 shows an example in which the number of the categories is 6. Apparently, this result is affected by the initial condition.

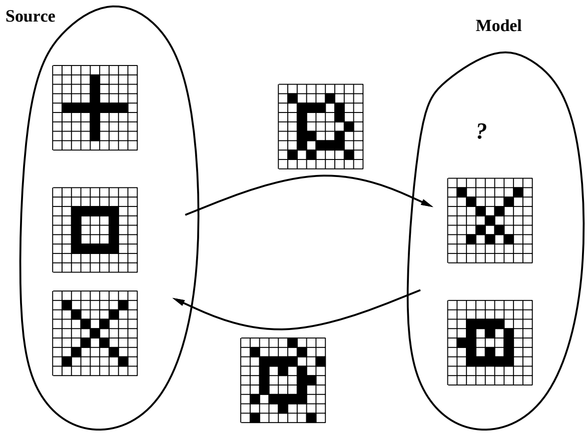
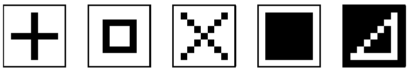


Fig. 3. The problem



error rate	.12	.11	.14	.17	.23
ξ_i	.22	.17	.18	.25	.18

Fig. 4. The Source

In most cases, one of the six categories is diminished into one of the other categories. Otherwise, the results are something like the figure that one of the categories is constructed for a few specific data, and it is corresponding to a category which doesn't exist in the source.

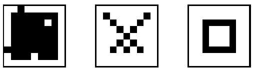
The procedure to select the categories is shown below.

1. Choose z to ask the source at random, according to the mixing weights π_z.
2. Generate x according to the density function $p_z(x|\theta_z)$, and ask it to the source. Then receive $\hat{x}$ from the source.
3. Repeat 1 and 2 k times and calculate (13), using (12).

$$s = \frac{1}{k}\sum_{s=1}^{k}\log p(\hat{x_s}, x_s|\theta) \qquad (13)$$

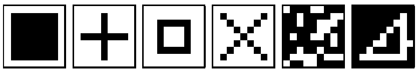
In the simulation, k is set to be 100. The results are shown in Table 1. According to the results, the model with 5 categories is selected because the result of it is the smallest. The model is shown in Figure 7.

Figure 7 shows that the selected model consists of correct number of categories and almost correct patterns.



error rate	.33	.22	.10
π_z	.36	.39	.25

Fig. 5. Mixture model with 3 mixture

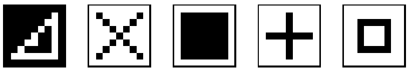


error rate	.15	.11	.10	.14	.19	.22
π_z	.20	.17	.25	.22	.02	.13

Fig. 6. Mixture model with 6 mixture

Categories	s
1	102.1870
2	91.7141
3	78.2310
4	72.7628
5	65.4223
6	65.5311
source	66.3330

Table 1. The value s of each model



error rate	.23	.14	.15	.11	.10
π_z	.16	.22	.20	.17	.25

Fig. 7. Final Model

5. Conclusion

In this article we proposed a method to construct the model by exchanging data with the source. In real world, there are many problems in which there must exist some categories but we cannot know it directly. We have given a learning algorithm for such problems and made a good result.

We are now thinking of applying this algorithm to more complicated problems. And also, we have to give statistical analysis to this algorithm.

References

[1] H. Akaike, "A new look at the statistical model identification," *IEEE Trans. Automat. Contr.*, vol. AC-19, pp. 716–723, Dec. 1974.

[2] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. R. Statistical Society, Series B*, vol. 39, pp. 1–38, 1977.

[3] D. M. Titterington, A. F. M. Smith, and U. E. Makov, *Statistical Analysis of Finite Mixture Distributions.* John Wiley & Sons, 1985.