On Approximation and Smoothing of General Distribution Based on Normal Expansion Method

Yang Jianhua* Student Member

Yasutaka Fujimoto* Member

A normal expansion method combined with MDL (Minimum Description Length) principle is proposed to construct an approximation of general distribution using its discrete data. Each of data is transformed to be a basic normal distribution and their composition is used to approximate the original distribution. Theoretically the composition is the same as the original distribution if the number of discrete data is infinite. For the finite quantity of data, a suitable approximation can be done if a variance is selected for each basic normal distribution to reach the MDL value. In this paper, the original distribution with single peak such as normal distribution, beta distribution etc., is firstly investigated and generated results show its feasibility and effectiveness. Then the original distribution with multiple peaks is approximated where the difference between use of identical variance and use of variant variance for each peak is also discussed. Finally as two applications, the smoothing of histogram and evaluation of data quantity are introduced to demonstrate utilization of proposed method.

Keywords : distribution approximation, distribution smoothing, normal expansion, MDL principle

1. Introduction

Probability model is applied to illustrate uncertainties of a system, predicting or evaluating the possible action of the system in the future ⁽¹⁾. However, it is often a hard problem to determine which kind of probability model should be introduced to describe data collected from a practical system. Generally there are three kinds of probability models, i.e., parametric model, non-parametric model and semi-parametric model ⁽²⁾. The type information of probability model, such as normal distribution, beta distribution, etc., should be known if parametric model is used. For parametric model, maximum likelihood method and Bayesian method are frequently employed to search optimal parameters ⁽³⁾⁽⁴⁾. The non-parametric model is applied if we know nothing about distribution type. A well-known method for building non-parametric model is histogram where kernel density method and K-nearest-neighbors method are often utilized to get a smoothing and size-suitable solution ⁽⁵⁾⁽⁶⁾. In this paper, a normal expansion method, which differs from kernel density method, is proposed to generate a smoothing and approximate distribution directly from sampling data. In kernel density method based on Parzen window ⁽⁷⁾, the statistical space should divided into many tiny domain and sufficient data should be provided. However it is unnecessary to divide statistical space in proposed method in this paper. The original distribution is rebuilt directly from sampling data and approximate solution can be obtained even if inadequate data are provided.

In normal expansion method, to rebuild the original distribution from sampling data, each datum is transformed to be a basic normal distribution and a combination of these basic normal distributions is regarded as approximation of original distribution. Theoretically, it can be proved that rebuilt distribution is the same

as the original one if we have infinite quantity of data. In practice, it is impossible to obtain infinite data by sampling therefore definite quantities of data are used to get an approximate solution by manipulating the variance of each basic normal distribution. To determine the suitable variance value, minimum description length (MDL) principle ⁽⁸⁾ is introduced, where the approximate distribution is regarded as a combination of generalized distributions subject to central limit theorem and the minimum description length is calculated using the number of generalized distributions. Furthermore, in this paper, identical variance and variant variance policies are adopted to investigate the accuracy of numerical results.

2. Normal Expansion Method

Let original probability distribution be $G(x)$ and its corresponding probability density function $g(x)$. The question is to rebuild original probability density function from N sampling data generated by $g(x)$. At first, y_i is supposed to be i -th sampled sampling data, subject to $y_1 \leq y_2 \leq \dots \leq y_{N-1} \leq y_N$. Then each of them is transformed to be a corresponding basic normal probability density function described by

$$q(x, y_i, \sigma, N) = \frac{1}{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x - y_i)^2}{2\sigma^2}\right) \dots\dots\dots (1)$$

where σ is a free parameter, corresponding to standard deviation of the basic normal distribution. Finally the combination of all basic normal probability functions can be represented by

$$\begin{aligned} f(x, \sigma, N) &= \sum_{i=1}^N q(x, y_i, \sigma, N) \\ &= \frac{1}{N} \sum_{i=1}^N \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x - y_i)^2}{2\sigma^2}\right) \dots\dots\dots (2) \end{aligned}$$

Based upon above definitions, we can obtain following theorem.

* Dept. of Electrical & Computer Engineering, Yokohama National University
79-5 Tokiwadai, Hodogayaku, Yokohama 240-8501

[Theorem]

$$\lim_{\delta \rightarrow 0} \lim_{N \rightarrow \infty} f(x, \sigma, N) = g(x) \quad (3)$$

[Proof]

Firstly a new variable $\omega = i/N$ subject to $0 < \omega \leq 1$ is introduced. According to definition of integral calculus, we have

$$\begin{aligned} \lim_{N \rightarrow \infty} f(x, \sigma, N) &= \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{i=1}^N \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-y_i)^2}{2\sigma^2}\right) \\ &= \int_0^1 \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-y(\omega))^2}{2\sigma^2}\right) d\omega \quad (4) \end{aligned}$$

For a large N , the i -th sampling data is situated at i/N over probability distribution $G(y)$ therefore we have

$$\frac{i}{N} \approx G(y) \quad (5)$$

When $N \rightarrow \infty$, it becomes

$$\omega = G(y) \quad (6)$$

whose differential results

$$d\omega = g(y)dy \quad (7)$$

Thus (4) can be rewritten as

$$\begin{aligned} \lim_{N \rightarrow \infty} f(x, \sigma, N) &= \int_0^1 \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-y(\omega))^2}{2\sigma^2}\right) d\omega \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-y)^2}{2\sigma^2}\right) g(y) dy \quad (8) \end{aligned}$$

At last

$$\begin{aligned} \lim_{\delta \rightarrow 0} \lim_{N \rightarrow \infty} f(x, \sigma, N) &= \lim_{\delta \rightarrow 0} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-y)^2}{2\sigma^2}\right) g(y) dy \\ &= \int_{-\infty}^{\infty} \lim_{\delta \rightarrow 0} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-y)^2}{2\sigma^2}\right) g(y) dy \\ &= \int_{-\infty}^{\infty} \delta(x-y) g(y) dy \\ &= g(x) \quad (9) \end{aligned}$$

where $\delta(x)$ is an impulse function.

Q.E.D

As pointed out in previous section, the case of infinite quantities of data is impossible in applications therefore suitable standard deviation σ should be determined for the case of finite quantity of data. At first, let's investigate the likelihood of approximate function, given by

$$L(\sigma) = \frac{1}{N} \sum_{i=1}^N \log f(y_i, \sigma, N) \quad (10)$$

It is obvious that likelihood will be maximized if $\sigma \rightarrow 0$. It implies that the distribution of $\sigma \rightarrow 0$ can only describe the sampling data themselves not the overall distribution. That is, at a point, it cannot receive any information from other sampling data.

Thus an additional condition should be introduced to get an approximate distribution.

MDL principle tells us that not only the complexity of the model but also the complexity of description of model itself should be taken in account. Based on MDL principle, the description length of approximate distribution can be computed by

$$D(\sigma) = -L(\sigma) + \frac{m}{2N} \log N \quad (11)$$

Herein, approximate probability function is regarded as a combination of peaks, each of which is supposed to be a generalized probability density function $p(\mu_k, \sigma)$ subject to central limit theorem. Assume that the dimension m is equal to the number of peaks. Consequently, approximate probability density function can be revised to be

$$f(x, \sigma, N) = \sum_{k=1}^m p(\mu_k, \sigma) \quad (12)$$

Apparently, $m \rightarrow N$ if $\sigma \rightarrow 0$, $m \rightarrow 1$ if $\sigma \rightarrow \infty$. And we have following results

$$\begin{aligned} D(\sigma)|_{\sigma \rightarrow 0} &= -L(\sigma)|_{\sigma \rightarrow 0} + \frac{N}{2N} \log N \\ &= -\frac{1}{N} \sum_{i=1}^N \log\left(\frac{1}{N} \sum_{j=1}^N \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_j - y_i)^2}{2\sigma^2}\right)\right)|_{\sigma \rightarrow 0} \\ &\quad + \frac{1}{2} \log N \\ &= -\infty \quad (13) \end{aligned}$$

$$\begin{aligned} D(\sigma)|_{\sigma \rightarrow \infty} &= -L(\sigma)|_{\sigma \rightarrow \infty} + \frac{1}{2N} \log N \\ &= -\frac{1}{N} \sum_{i=1}^N \log\left(\frac{1}{N} \sum_{j=1}^N \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_j - y_i)^2}{2\sigma^2}\right)\right)|_{\sigma \rightarrow \infty} \\ &\quad + \frac{1}{2N} \log N \\ &= +\infty \quad (14) \end{aligned}$$

It implies that traditional MDL principle cannot give us a solution thus a modification should added. The suitable standard deviation σ should be determined except 0 for finite data because no information can be received from other data at $\sigma=0$ as described previously. Moreover the parameter m is a discrete variable therefore following formula is employed to obtain a proper solution.

$$\begin{cases} D(\sigma - d\sigma) > D(\sigma) \\ D(\sigma + d\sigma) > D(\sigma) \\ \sigma > 0 \end{cases} \quad (15)$$

Usually, there exist multiple solutions subject to (15) and the local minimum one is selected.

3. Normal Expansion Method

To verify our proposed normal expansion method, a lot of parametric models and uniform distribution, an element of nonparametric model, are used to generate sampling data.

3.1 Single Peak Distribution Many probability density functions are of single peak type. In this paper, normal distribution,

beta distribution, 2-order function distribution and uniform distribution are investigated.

At first, normal distribution is considered and its probability density function is represented by

$$g(x) = \frac{1}{\sqrt{2\pi}\sigma_0} \exp\left(-\frac{(x-\mu)^2}{2\sigma_0^2}\right) \dots\dots\dots (16)$$

As an example of computation, let σ_0 be 10, μ an arbitrary value because the average μ is independent of approximating problem. Let the number of sampling data N be 20 and the distance among data is subject to probability density function of normal distribution $g(x)$. The approximating procedure using previously described approach is depicted in Fig.1. Fig.1 shows that 20 peaks exist when σ is very small and the number of peaks decreases to be one finally when σ gradually increases. It is also known that the relationship among data cannot be represented when $\sigma=0.2$ even if its corresponding MDL value is very small, and the difference among data cannot be represented as well when $\sigma=5$, an exceeded value. The final selected solution is $\sigma=1.8$ using the condition given in (15) in previous

(a) Approximation Procedure of Normal Probability density function ($\sigma_0 = 10$) When $N = 20$

(b) The Number of Peaks v.s. σ

(c) MDL Value v.s. σ

Fig. 1. Approximation of Normal Probability density function ($\sigma_0 = 10$) When $N = 20$.

section.

Similarly, let the number of data N be 10,20,30,40,50,60,70, 80,90,100 and investigate their accuracies. The error between approximate probability density function and original probability density function is evaluated by

$$e = \frac{\int |g(x) - f(x, \sigma, N)| dx}{\int g(x) dx} \times 100\% \\ = \int |g(x) - f(x, \sigma, N)| dx \times 100\% \dots\dots\dots (17)$$

Numerical results are shown in Table 1, among which the approximate probability density functions of $N = 10, 20$ are shown in Fig.2

Above results come to a conclusion that the approximate probability density function closely resembles original one and the standard deviation σ becomes small when the number of data increases.

Next, the second example is beta distribution whose probability density function is represented by

$$g(x) = \frac{B(\alpha, \gamma)}{(b-a)^{\alpha+\gamma+1}} (x-a)^\alpha (b-x)^\gamma \quad (a < x < b) \dots\dots\dots (18)$$

$$g(x) = 0 \quad (x \leq a, x \geq b) \dots\dots\dots (19)$$

$$B(\alpha, \gamma) = \frac{\Gamma(\alpha + \gamma + 2)}{\Gamma(\alpha + 1)\Gamma(\gamma + 1)} \dots\dots\dots (20)$$

$$\Gamma(\lambda) = \int_0^\infty x^{\lambda-1} e^{-x} dx \dots\dots\dots (21)$$

where $\Gamma(\lambda)$ is gamma function and we have $\Gamma(\lambda) = (\lambda - 1)!$ if λ is an integer.

For beta distribution, four parameters can be reduced to be three ones if we know the length of interval $(b - a)$. As an example of computation, let α be 1, γ be 4 and $(b - a)$ be 80, then corresponding beta distribution becomes

Table 1. Error Vs data quantity (Normal Distribution).

N	10	20	30	40	50
σ	2.0	1.8	1.6	1.5	1.4
e [%]	18.57	9.72	6.79	5.20	4.26
N	60	70	80	90	100
σ	1.4	1.4	1.4	1.4	1.4
e [%]	3.52	3.00	2.61	2.30	2.06

Fig. 2. Approximation of Normal Probability density function ($\sigma_0 = 10$).

Table 2. Error Vs data quantity (Beta Distribution).

N	10	20	30	40	50
σ	3.2	2.8	2.4	2.2	2.2
e [%]	14.92	7.75	5.40	4.15	3.39
N	60	70	80	90	100
σ	2.0	2.0	2.0	2.0	1.8
e [%]	2.89	2.56	2.33	2.18	1.92

Fig. 3. Approximation of beta probability density function.

Table 3. Error Vs data quantity (2-order Function Distribution).

N	10	20	30	40	50
σ	1.7	1.2	0.9	0.8	0.7
e [%]	16.46	8.82	6.26	4.68	3.81
N	60	70	80	90	100
σ	0.7	0.6	0.6	0.5	0.5
e [%]	3.05	2.73	2.31	2.20	1.92

Fig. 4. Approximation of 2-order function distribution.

$$g(x) = \frac{30}{80^6} (x-a)(b-x)^4 \quad (22)$$

Let the number of data N be 10,20,30,40,50,60,70,80,90, 100. Numerical results are obtained and shown in Table 2, among which the approximate probability density functions of $N = 10, 20$ are shown in Fig.3

Next, the third example is 2-order function distribution whose probability density function is represented by

$$g(x) = (x-a)(x-a+40)/10666.6 \quad (23)$$

where a is arbitrary.

Let the number of data N be 10,20,30,40,50,60,70,80,90, 100. Numerical results are obtained and shown in Table 3, among which the approximate probability density functions of $N = 10, 20$ are shown in Fig.4.

Finally uniform distribution is considered, whose probability density function is represented by

Fig. 5. Approximation of uniform distribution ($b-a=40$) When $N=10$ and $\sigma=4.4$.

Table 4. Error Vs data quantity (Mixed Normal Distribution).

N	10	20	30	40	50
σ	4.9	2.0	1.9	1.9	1.1
e [%]	21.52	7.39	5.82	5.26	3.79
N	60	70	80	90	100
σ	1.9	1.9	1.9	1.9	0.8
e [%]	4.90	4.87	4.86	4.88	2.19

$$g(x) = \begin{cases} \frac{1}{b-a} & a < x < b \\ 0 & x \leq a, x \geq b \end{cases} \quad (24)$$

Uniform distribution is an exception of single peak distribution and the basic component of histogram as well. In Uniform distribution, the density within intervals $x \leq a$ and $x \geq b$ equals 0 therefore the part of $x \leq a$ and $x \geq b$ should be compensated into valid interval $a < x < b$. Let's look at an example when $b-a=40$ and $N=10$, whose numerical result without compensation is shown in Fig.5(a). Fig.5(b)(c) illustrates a boundary reflex method, with which the outside part will be reflected into inside part at the bound point one time. The boundary reflex procedure can be described as follows.

$$\tilde{f}(x, \sigma, N) = \begin{cases} f(x, \sigma, N) + f(2b-x, \sigma, N) & a < x < b \\ + f(x-2a, \sigma, N) & \\ 0 & x \leq a, x \geq b \end{cases}$$

where $\tilde{f}(x, \sigma, N)$ is the reflected result. And compensated result using boundary reflex is shown in Fig.5(d) where the error has been greatly reduced.

Generally boundary reflex compensation should be carried out for a single peak with specified bound if the probability density nearby the bound is larger than 0.

3.2 Multi-Peak Distribution The first example of multi-peak distribution is a combination of normal distribution, whose probability density function is represented by

$$g(x) = \left(\frac{1}{\sqrt{2\pi}\sigma_a} \exp\left(-\frac{(x-\mu_1)^2}{2\sigma_a^2}\right) + \frac{1}{\sqrt{2\pi}\sigma_b} \exp\left(\frac{(x-\mu_2)^2}{2\sigma_b^2}\right) + \frac{1}{\sqrt{2\pi}\sigma_c} \exp\left(-\frac{(x-\mu_3)^2}{2\sigma_c^2}\right) \right) \times \frac{1}{3} \quad (25)$$

Let $\sigma_a = 4$, $\sigma_b = 10$, $\sigma_c = 5$, $\mu_1 = \text{arbitrary}$, $\mu_2 = \mu_1 + 10$, $\mu_3 = \mu_2 + 10$. The errors of approximate distribution are shown in Table 4 when $N = 10, 20, 30, 40, 50, 60, 70, 80, 90, 100$, among which the approximate probability density functions of $N = 10, 20$ are shown in Fig. 6

The second example of multi-peak distribution is based on a following complexity probability density function

Fig. 6. Approximation of mixed normal distribution.

Table 5. Error Vs data quantity (Complex Distribution).

N	10	20	30	40	50
σ	5.0	1.4	1.0	0.8	0.6
e [%]	23.76	8.39	5.65	4.27	3.45
N	60	70	80	90	100
σ	0.6	0.5	0.4	0.4	0.4
e [%]	2.89	2.48	2.21	1.94	1.75

Fig. 7. Approximation of complex distribution.

Fig. 8. Approximation ($\delta = 5.0$) of distribution considering bound when $N = 10$.

$$g(x) = (1 + \sin(0.02(10 + 0.1x + 0.08x + 0.001x^3 + 0.0002x^4 + 0.00002x^5))) / 52.74 \quad (26)$$

$(-20 < x < 20)$

The numerical results of approximate distribution are shown in Table 5 when $N = 10, 20, 30, 40, 50, 60, 70, 80, 90, 100$, among which the approximate probability density functions of $N = 10, 20$ are shown in Fig. 7

The probability density function given in (26) clearly specifies the bound 20 and -20. The approximating result can be improved if boundary compensation is added. Fig. 8 shows its improved result where error has been reduced by 7.39%.

From the results described in section 3.1 and 3.2, following conclusion can be induced. Generally a probability density function can be rebuilt using proposed normal expansion method and satisfied approximation can be provided if sufficient sampling data are given.

4. Investigation on Variance

Previous results show that errors are often serious if a small quantity of data is provided. Following is an example in which approximate distribution cannot rebuild original distribution properly. Let $\sigma_a = 2$ in (28). Using identical variance, the approximate results are depicted in Fig. 9 when $N = 10, 20$. The problem is that only the peak, $\sigma_a = 2$, can be rebuilt for $N = 10$. If we return to the graph of MDL curve and investigate the approximating procedure given in Fig. 10, we find that multiple solutions of (15) might have their practical meanings. It seems that some of solutions can be also used to approximate corresponding peaks. In fact, Fig. 11 demonstrates that the peaks, $\sigma_c = 5$ and $\sigma_b = 10$, can be rebuilt using the solutions, $\sigma = 2.2$ and $\sigma = 3.5$, respectively.

So far we have been employing identical variance to rebuild original distribution. Above phenomenon implies that it seems that variant variance is a more suitable solution to solve above problem. If different variance is specified for each data, formula (1) will be modified to be

$$q(x, y_i, \sigma_i, N) = \frac{1}{N} \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(-\frac{(x-y_i)^2}{2\sigma_i^2}\right) \quad (27)$$

Although pure variant variance might rebuild original distribution more reasonably, an unavoidable reality is that it is hard to determine all variant variances for all data. If it is supposed that the same variance is utilized for the data on the

Fig. 9. Approximation of Mixed Normal Distribution ($\sigma_a = 2$).

same peak, the question becomes simple and we just need to know which data are situated on corresponding peak.

Following algorithm is developed based on above idea that peaks are searched one by one.

Step 1: An initial identical variance is employed for all data to obtain the solution subject to (15), minimizing the MDL value. Let the solution be σ_1 .

Step 2: Multiple peaks might exist. Count the data for each peak

(a) The Number of Peaks v.s. σ

(b) MDL Value v.s. σ

Fig. 10. Approximation of mixed normal distribution ($\sigma_a = 2$) When $N = 10$.

Fig. 11. Approximation of mixed normal distribution ($\sigma_a = 2$) When $N = 10$.

Fig. 12. Approximation of mixed normal distribution ($\sigma_a = 2$) Using Variant Variance.

and the standard deviation is fixed to be σ_1 for the data on the peak with most data. Define the peak with most data be $p(\mu_1, \sigma_1)$.

Step 3: For the data out of the peak $p(\mu_1, \sigma_1)$, identical variance is employed once again. Let the current solution be σ_2 and count the data for each peak except $p(\mu_1, \sigma_1)$. Define the current peak with most data be $p(\mu_2, \sigma_2)$. The standard deviation is fixed to be σ_2 for the data on the peak $p(\mu_2, \sigma_2)$.

Step 4: Above procedure continues until no datum is left.

The result given in Fig.9(a) is enhanced using above algorithm. Fig.12 shows its improved result where two peaks appear. The reason why the peaks, $\sigma_c = 5$ and $\sigma_b = 10$, are rebuilt using the same standard deviation 4.8 is because both of them are similar and the quantity of data is insufficient.

In Fig.12(b), 4 iterations are performed but 3 peaks are formed because the last one is absorbed into another one. The Fig.12(c)(d) demonstrate that identical variance based method is also a possible choice if sufficient data are provided.

5. Smoothing Histogram

Proposed normal expansion method shows that more data provided will rebuild original distribution more accurately. However, too many sampling data will result in time consuming and model explosion. Therefore smoothing histogram generated from a large quantity of data is a possible solution.

Fig.13 shows traffic density per hour of a road during a day. The density is represented by the number of cars passing through the road per hour divided by the total of cars passing through during a day. It is supposed that the signal interchange should be determined by the traffic density and the time interval should be smoothly changed. That is, smoothing the histogram shown in Fig.13 is necessary.

Histogram can be regarded as another representation of probability density function, with which re-sampling can be performed. Let the distances among data obey the histogram,

Fig. 13. Histogram of traffic density per hour.

Fig. 14. Smoothing of histogram of traffic density.

Fig. 15. Smoothing of histogram of traffic density using frequency distribution.

generating a series of data. And Fig.14 shows the results of smoothing histogram when the number of data $N = 10, 20$ based on variant variance method.

The errors given in Fig.14 is not the errors between original distribution and approximate distribution but the errors between histogram and approximate distribution due to unknown original distribution. The errors are a little larger than that of approximating general distribution with smooth curve because of histogram characteristics. If more data are provided, the approximation will be more closely similar to histogram.

On the other hand, another simpler method can be applied to solve the problem of smoothing histogram, where the number of data falling in each interval is introduced. That is, the approximate probability density function can be represented by

$$f(x, \sigma, N) = \sum_{j=1}^M \hat{g}(y_j) \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-y_j)^2}{2\sigma^2}\right) \dots\dots\dots (28)$$

where $\hat{g}()$ is the histogram. And the MDL value is calculated by

$$\begin{aligned} D(\delta) &= -\frac{1}{N} \sum_{i=1}^M N \hat{g}(y_i) \log f(y_i, \delta, N) + \frac{m}{2N} \log N \\ &= -\sum_{i=1}^M \hat{g}(y_i) \log f(y_i, \delta, N) + \frac{m}{2N} \log N \dots\dots\dots (29) \end{aligned}$$

For the problem described in Fig.13, $M = 24$ is fixed. Fig.15 shows the approximate results when $N = 10, 20$ using (29). It demonstrates that the errors are independent of N if N is a sufficient large number. In fact, the procedure of approximating procedure in Fig.15 is always based on 24 normal probability density functions. Just a sufficient large value is needed for N because N is just a computation factor. It is very the advantage of the (28) where computation time is fixed.

6. Evaluation of Information Amount

Entropy is often applied to evaluate the information amount among data. It can be obtained based on probability density function if continuous variable function is employed. Hereby the question is that only discrete data are provided and continuous probability density function should be firstly constructed based on those data.

For example, one-machine scheduling problem is considered. It is well known that SPT (shortest processing time) rule is an optimal solution to minimization of average completion time of jobs. Nevertheless, arbitrary sequence is allowed if all jobs are

Fig. 16. Evaluation of similarity for Jobs.

with the same processing time. It implies that the amount of difference among jobs might reveal the importance of scheduling algorithm. In other words, the higher similarity among jobs is, the less the importance of scheduling will be. The similarity among jobs can be evaluated by entropy therefore the importance of scheduling algorithm can be expressed by obtained entropy as well.

Supposed that there are 20 jobs waiting in buffer before the machine and the processing time of jobs can vary between 10 and 50 minutes. Four cases with different processing time are given as follows.

$J_1 = \{31, 45, 37, 37, 31, 35, 23, 44, 40, 19, 43, 34, 43, 35, 14, 38, 23, 27, 27, 31\}$,
 $J_2 = \{19, 30, 42, 44, 23, 40, 46, 32, 25, 21, 41, 25, 15, 47, 26, 21, 49, 39, 42, 27\}$,
 $J_3 = \{12, 46, 46, 26, 28, 26, 44, 22, 35, 14, 49, 16, 37, 47, 30, 24, 12, 44, 33, 11\}$,
 $J_s = \{11, 13, 15, 17, 19, 21, 23, 25, 27, 29, 31, 33, 35, 37, 39, 41, 43, 45, 47, 49\}$.

Fig.16 depicts all continuous probability density functions for all cases where variant variance based method is employed and boundary reflex compensations are performed for all results.

Fig.16 clearly demonstrates that the job with higher entropy is of lower similarity. The extreme case is shown in Fig.16(d) where jobs are with uniform processing time thus they are the most different each other and a scheduling algorithm might have the most significant effect on the objective function.

7. Conclusions

In this paper normal expansion method combined with MDL principle is proposed to rebuild an approximation of probability density function based on discrete sampling data generated from original distribution. In proposed method, each datum is transformed into a basic normal probability density function and

the approximate probability density function is a combination of those normal probability density functions. Based on theoretical analysis and numerical result for various distributions, we have following conclusions.

- (1) It is proved that original distribution can be completely restored using normal expansion method if infinite data are provided. It be also demonstrated that original distribution can be approximately rebuilt by applying MDL principle to search a suitable variance if finite data are given.
- (2) For single peak and multi-peak distributions, approximating errors decrease if the number of data increases.
- (3) For multi-peak distributions, the difference between use of identical variance and use of variant variance is investigated. Identical variance based method is also possible solution if sufficient data are provided although variant variance based method is a general one.
- (4) smoothing histogram and evaluating information amount are introduced to demonstrate the possible practical use.

(Manuscript received Oct. 27, 2003, revised March 15, 2004)

References

- (1) D. P. Heyman : M. J. Sobel: Stochastic Models, Elsevier Science (1990)
- (2) C. M. Bishop : Neural Networks for Pattern Recognition, Oxford Univ. Press (1995)
- (3) A. P. Dempster, N. M. Laird, and D. B. Rubin : "Maximum Likelihood from Incomplete Data via The EM Algorithm", *Journal of Royal Statistical Society*, Vol.39, pp.1-38 (1977)
- (4) J. Whitehead : "Bayesian decision procedures with application to dose-finding studies", *International Journal of Pharmaceutical Medicine*, Vol.11, pp.201-208 (1997)
- (5) G. Terrell and D. Scott : "Variable Kernel Density Estimation", *Ann. Statistic*, Vol.20, No.3, pp.1236-1265 (1992)
- (6) T. K. Ho : "Nearest Neighbors in Random Subspaces", Lecture Notes in Computer Science: Advances in Pattern Recognition, pp.640-648 (1998)
- (7) E. Parzen : "On Estimation of A Probability Density Function and Mode", *Annals of Mathematical Statistics*, Vol.33, pp.1065-1076 (1962)
- (8) J. Rissanen : "Modeling by Shortest Data Description", *Automatica*, Vol.14, pp.465-471 (1978)

Yang Jianhua

(Student Member) was born in Jiangsu province, China, in 1968. He received the B.S. degree from South-east University, Nanjing, China in 1990 and the M.S. degree from Tsinghua University, Beijing, China in 1993. He had worked in Automation Department of Tsinghua University from 1993 to 2000. He has been a doctoral candidate of the Department of Electrical and Computer Engineering, Yokohama National University since 2001. His research interests include flexible manufacturing system, shop-floor control and Petri Nets. Mr. Yang is a student member of the Institute of Electrical Engineers of Japan.

Yasutaka Fujimoto

(Member) was born in Kanagawa prefecture, Japan, in 1971. He received B.E., M.E., and Ph.D. degrees in electrical engineering from Yokohama National University, Japan, in 1993, 1995, and 1998, respectively. In 1998, he joined the Department of Electrical Engineering, Keio University. Since 1999, he has been with the Department of Electrical and Computer Engineering, Yokohama National University, where he is currently an Associate Professor. His research interests include manufacturing automation, discrete event systems, motion control, and robotics. He is a member of IEEE and Robotics Society of Japan.