42

HUMAN FACTOR AND ITS IDENTIFICATION IN MANUFACTURING PREDICTION

Yang Jianhua and Yasutaka Fujimoto
Dept. of Electrical and Computer Engineering, Yokohama National Univ. Japan
e-mail: yangjh@fujilab.dnj.ynu.ac.jp, fujimoto@ynu.ac.jp

Abstract: A decision model, stemmed from Bayesian thinking, is proposed to predict the operator's behavior in manufacturing system. The decision model is addressed using non-parametric distribution where a binary division method is proposed to reduce the complexity of the model, eliminating irrelevant features.

Key words: human factor, prediction, Bayesian theorem, non-parametric model

1. INTRODUCTION

In general a manufacturing plan should be set up to meet the time constrains of orders, while the delivery dates are also determined by prediction of complete time of orders based on the capacity of manufacturing system. In many cases, we find that a shop floor is controlled by a group of operators who make their decisions according to some rules usually given by a guide, their experience obtained from manufacturing history, and probably their mood of that day. Therefore, the human interference should be included if we hope to predict the future of manufacturing. Errors might be decreased if we can correctly identify the behaviors of operators using the past decision data.

The identification of human behavior in manufacturing system differs from general pattern recognition[1] in which sampling data, the manufacturing history, have been given and random sampling is not applicable. Therefore sufficient manufacturing history data are needed to make it possible to identify the human behavior. The model used in this

paper is primarily developed and originated from Bayesian thinking, where some special transformations are introduced for constituting feature vector from parameters of factors. On the other hand, a full non-parametric model is proposed to dispose of both continuous and discrete variables with irregular distributions. To solve the complexity problem where a non-parametric model might result in an explosion of data storage[2] and the relevance selection problem where irrelevant factors are redundant[3], we propose a binary division method in this paper.

2. MODELLING

With respect to the problem we consider, it is unnecessary to obtain a general model of human's brain. In fact, it is also unrealistic until today even if we have received some clues about its operation mechanism. In order to clearly illustrate the essence of the problem, we formulate the process of human's decision making as follows. Let the operator's surroundings be C, the operator's status be M, the operator's decision mechanism be D, and the final decision be Q. The process of human's decision making can be represented by $D: C \times M \rightarrow Q$. Unfortunately we often only know the partial surrounding information I around the operator and try to employ $D: I \rightarrow Q$ to represent the decision process of the manufacturing system. In other words, the decision recognition can be expressed by $R: (I,Q) \rightarrow D$, where R stands for recognition mechanism. Errors are inevitable if $(C-I) \times M \neq \phi$. Therefore the topic about error decreasing in this paper is always discussed over R.

Bayesian thinking[4] is often employed in pattern recognition. Let Ω be sampling space, which is composed of n independent hypotheses, noted by $\{B_1, B_2, \cdots, B_n\}$. The probability of occurrence for result x can be computed by following equation:

$$p(B_k \mid x) = p(x \mid B_k) p(B_k) / p(x) \tag{1}$$

For identification of operator's behavior, we let hypotheses be operator's decision D, let results be history system data I and history decision data Q. According to equation (1), we get

$$p(D \mid (I,Q)) = p((I,Q) \mid D) p(D) / p((I,Q)) \tag{2}$$

Furthermore, suppose that an operator always select no more than one job based on the current status of surroundings. The surroundings data at that time are addressed by a feature vector X, then (2) can be expressed by

$$p(D \mid X) = p(X \mid D) p(D) / p(X) \tag{3}$$

From manufacturing history, distribution $p(X \mid D)$ and $p(X)$ can be obtained although sometimes it is hard and complex to do so. For prior distribution $p(D)$, suppose that it is a uniform distribution. As a result, the

posterior distribution $p(D|X)$ is fundamentally determined by $p(X|D)/p(X)$. Let

$$\eta(X|D) = p(X|D)/p(X). \tag{4}$$

The $\eta(X|D)$ can be regarded as a force, the operator's decision, driving the prior probability distribution $p(X)$ to the posterior probability distribution $p(X|D)$. Finally for the future status, we predict that operator will select a job $j^* \in J$ such that $p(D|X^F_{j^*}) = \max_j p(D|X^F_j)$, i.e.,

$$\eta(X^F_{j^*}|D) = \max_j \eta(X^F_j|D) \tag{5}$$

3. DECISION ACQUISITION

As illustrated in Section 2.1, the decision recognition distribution $\eta(X|D)$ naturally describes the operator's decision mechanism and shows how much information is obtained.

To simply illustrate it, let $[\bullet]$ be logarithm of probability variable $p(\bullet)$, then we have

$$[X|D] = \log p(X|D),\ [X] = \log p(X). \tag{6}$$

And we define

$$[\eta(X|D)] = [X|D] - [X] \tag{7}$$

$$F(D) = \int |[X|D] - [X]|\, dX \tag{8}$$

where $F(D)$ is referred to as intensity of operator's decision and $\int$ stands for a generalized integral operator which can compute over both discrete and continuous variables.

In general, the higher the intensity means the stronger operator's decision. If the $p(X)$ have the same distribution with $p(X|D)$, we get $F(D) = 0$. It means that we learn nothing from manufacturing history, i.e., one might select jobs randomly. Therefore we certainly cannot predict the future. But for an effective identification, we always have $F(D) > 0$. For instance, only a feature x is considered. Given $p(x=0) = 0.6$ and $p(x=1) = 0.4$. After operator's decision, we get $p(x=0|D) = 0.8$ and $p(x=1|D) = 0.2$. Then how much can we learn from history? Or $F(D) = ?$.

Here the $\int$ is substituted by the $\sum$, then we get

$$F(D) = \sum |[X|D] - [X]|$$

$$= \sum |\log p(X|D) - \log p(X)| = |\log 0.8 - \log 0.6| + |\log 0.2 - \log 0.4| = 0.426$$

Note that definition of $\eta[X|D]$ is invalid if $p(X|D) = 0$ or $p(X) = 0$. Therefore integral $F(D)$ defined in formula (8) does not always exist. To strictly define it, we discuss some properties over so-called valid sampling space. Let Ω_R stand for valid sampling space of $p(X)$ such that $p(X) > 0$,

Ω_D for valid sampling space of $p(X \mid D)$ such that $p(X \mid D) > 0$, respectively. We have following conclusions.

[Property 1] $\Omega_R \geq \Omega_D$.

It can be obviously proved because a sampling point of the operator's decision should be one belonging to original sampling space. Particularly, $\Omega_R = \Omega_D$ means that no more than one job waits in the buffer at any time hence the operator has no choice but select the only one.

[Property 2] $p(X) > 0$ if $p(X \mid D) > 0$.

It can be induced from *property 1* and can be regarded as another description of *property 1*.

Based on *property 2*, we can revise formula (8) as

$$F(D) = \int_{\Omega_D} |[X \mid D] - [X]| dX \tag{9}$$

whose definition always exists.

[Property 3] let $\Delta\Omega = \Omega_R - \Omega_D$, we have

$$p(X) > 0 \text{ and } p(X \mid D) = 0 \text{ for } \forall X \in \Delta\Omega.$$

It can be concluded from *property 2* and the definition of valid sampling space. The domain $\Delta\Omega$ is also referred to as deterministic decision space, implying that a sampling point in $\Delta\Omega$, which is also referred to as deterministic decision point, will be surely recognized because zero is the smallest value. Generally the larger the domain $\Delta\Omega$ is, the stronger the decision mechanism is.

4. NON-PARAMETRIC DISTRIBUTION

The simplest way to describe distribution of $\eta(X \mid D)$ is utilization of parametric model such as normal distribution, beta distribution etc, where the distribution can be completely represented by some parameters such as average value, variance, etc. However the distribution type should be known before we employ parametric model. Thus to obtain general description of $\eta(X \mid D)$, non-parametric model is usually a possible choice.

A non-parametric distribution model is generally described by dividing sampling space into many tiny domains, where distribution density $p(X)$ is almost constant. Let a domain be S, corresponding volume be V. The probability of feature vector in S can be calculated by $p(S) \cong p(X)V$. According to Monte Carlo simulation[5], given m sampling data, if among them k data fall in the domain S, the probability of feature vector in S can be obtained by $p(S) = k/m$. Thus the distribution density in domain S can be determined by

$$p(X) = k/mV. \tag{10}$$

The basic two methods for modeling non-parametric distribution are kernel density method[6][7] and k-nearest neighbors method[8]. For the kernel density method, the probability density of a domain can be calculated

by fixing the volume of the domain, counting the data that fall in it. For k-nearest neighbor method, the probability density of a domain can be calculated by fixing the number of data that fall in the domain, changing the volume of the domain. A main drawback of the kernel density method is that a large domain division might result in low smooth while a small domain division might result in low reliability due to limited history data. Moreover, sometimes its implementation is almost infeasible. The k-nearest neighbors method emphasizes that the volume of domain is changeable, fixing the counts of data that fall in the domain. But it is often hard to get such a domain. In fact, no matter what kind of method, the fundamental problem is the division of sampling space. In next section, a binary division method is proposed to provide such a solution, where both the volume and counts are changeable.

5. BINARY DIVISION METHODOLOGY

Noticed that an effective decision means that decision distribution $p(D|X)$ is not a uniform distribution. The larger difference among domains generally implies the more effective decision. So we should emphasize the feature with less variance and consider how to divide it firstly. Here a binary division method is one of possible choices.

Let Ω be the sampling space, $X=[x_1 \ \ x_2 \ \cdots \ x_k]$ be a feature vector. At first, a binary division is done along each feature $x_i (i=1,2,\cdots,k)$, so we get a group of bi-subspaces, i.e., domains, denoted by $S(x_i,L,\Omega)$ and $S(x_i,R,\Omega)$, where L stands for the left domain, R for right domain, respectively. As described previously, instead of computing probability density $p(D|X)$, $\eta(X|D)$ is applied to describe recognition distribution therefore we define $\eta(x_i,L,\Omega|D)$ standing for density distribution of $S(x_i,L,\Omega)$, $\eta(x_i,R,\Omega|D)$ for density distribution of $S(x_i,R,\Omega)$. Among k divisions only one along the feature x_{i^*} ($i^*\in\{1,2,\cdots,k\}$) is really selected to be executed, which is such that

$$\Delta\eta(x_{i^*},\Omega)=\max_i \Delta\eta(x_i,\Omega) \tag{11}$$

where $\Delta\eta(x_i,\Omega)=|\eta(x_i,L,\Omega|D)-\eta(x_i,R,\Omega|D)|$. (12)

Similarly for each subspace $S_u\in\{S(x_{i^*},L,\Omega), S(x_{i^*}, R,\Omega)\}$ we can obtain its furthermore divided subspaces $S(x_i,L,S_u)$ and $S(x_i,R,S_u)$ by binary divisions. And the really executed division along the feature x_{i^*} ($i^*\in\{1,2,\cdots,k\}$) at this step is also such that

$$\Delta\eta(x_{i^*},S_u)=\max_i \Delta\eta(x_i,S_u), \tag{13}$$

where $\Delta\eta(x_i,S_u)=|\eta(x_i,L,S_u|D)-\eta(x_i,R,S_u|D)|$. (14)

Apparently such a division might be carried out infinitely, producing countless domains therefore a termination condition should be added.

Hereby, we introduce two thresholds: an integer $\sigma(\geq 0)$ standing for a threshold of sampling points for a subspace S_u and a real number $\delta(\geq 0)$ for a threshold of the difference of density distribution between two subspaces of the subspace S_u. The binary division process will be stopped if

$$C(S_u) \leq \sigma \,\|\, \Delta\eta(x_{i^*}, S_u) \leq \delta \tag{15}$$

where $C(S_u)$ is the sampling points of the subspace S_u and symbol $\|$ represents 'OR' Boolean operator.

The domain division for non-parametric distribution is equivalent to sampling problem in signal processing. An effective technique is that the higher density makes more divisions, vice versa. It is the threshold σ that determines how small a domain should be.

Furthermore, as we consider the problem of division of sampling space, distinguishing relevant and irrelevant features should be also taken in account. Clearly the model will become redundant if an irrelevant feature is involved. Therefore is it possible that irrelevant features can be kicked out when domains are divided?

It is clear that the times of binary division along the each feature x_i, denoted by $\kappa(x_i)$, might be different. And it can be applied to deal with the problem of elimination of irrelevant features. Before some conclusions are induced, the definitions of irrelevant feature are discussed as follows.

[Definition 1] Irrelevant feature in strong sense: A feature x_r is an irrelevant feature if decision distribution $p(D|x_r)$ is a uniform distribution and independent of other features.

Using above definition and the sampling division method, we obtain the following theorem.

[Theorem] The times of binary division along a feature x_r is denoted by $\kappa(x_r)$. $\kappa(x_r) = 0$ if the feature x_r is irrelevant to operator's decision in strong sense.

[Proof]

Based on equations (3), (4), we get

$$\eta(X|D) = p(X|D)/p(X) = p(D|X)/p(D). \tag{16}$$

For the feature x_r, we have

$$\eta(x_r|D) = p(D|x_r)/p(D). \tag{17}$$

The distribution $\eta(x_r|D)$ should be uniform because $p(D|x_r)$ and $p(D)$ are uniform distributions, according to definition and assumption.

The uniform property is kept for all domains if a feature is independent of others, therefore the binary division on x_r for any subspace S_u is always such that

$$\Delta\eta(x_r, S_u) = 0. \tag{18}$$

But according to binary division method, only the binary division such that $\Delta\eta(x_i, S_u) > 0$ is possibly selected and really executed. Thus the binary

division will be never really executed on x_r, i.e.,

$$\kappa(x_r) = 0. \tag{19}$$

[End]

However we cannot induce that a feature is irrelevant one in strong sense even if $\kappa(x_r) = 0$ using proposed binary division. Therefore we introduce the definition of irrelevant feature in weak sense as follows.

[Definition 2] Irrelevant feature in weak sense: A feature x_r is an irrelevant one if $\kappa(x_r) = 0$.

That is, we can eliminate irrelevant features in weak sense using binary division method.

6. AN EXAMPLE

Given a set of jobs J = {1,2,3,4,5,6,7,8,9,10}, waiting in a buffer to be processed on a machine, its corresponding processing time and parts size are represented by $H = \{h(j)\}_{10}$ = {12,24,30,28,48,51,61,60,70,66} and $S = \{s(j)\}_{10}$ = {20,10,10,20,30,30,30,20,20,10} respectively. Let the parts size of the job before job 1 be 20. Suppose that jobs are mounted according to sequence $1 \rightarrow 2 \rightarrow 3 \rightarrow 4 \rightarrow 5 \rightarrow 6 \rightarrow 7 \rightarrow 8 \rightarrow 9 \rightarrow 10$. Define feature vector $X = [x_1 \quad x_2]^T$ where

$$x_1 = h(i) - \min_j h(j) \quad \text{(Job } i \in J \text{ is the next one to be mounted, } j \neq i\text{)} \tag{20}$$

$$x_2 = \begin{cases} 0 & s(o) = s(i) \\ 1 & s(o) \neq s(i) \end{cases} \text{ (Job } o \in J \text{ is the one that just has been processed).} \tag{21}$$

To simplify our example, we suppose that $p(X)$ is approximately a uniform distribution. Therefore $\eta(X \mid D)$ is determined by $p(X \mid D)$. According to above processing sequence, we obtain a set of sampling data

$$\left\{ \begin{matrix} -12 & -4 & 2 & -24 & -3 & -9 & 1 & -6 & 4 \\ 0 & 1 & 0 & 1 & 1 & 0 & 0 & 1 & 0 \end{matrix} \right\},$$

where we needn't make a decision for the last job thus only 9 data are generated. Let $\sigma = 1, \delta = 0$. The result of binary division along $X = [x_1 \quad x_2]^T$ is shown in *Fig.1* and obtained histogram is shown in *Fig.2*.

Fig.1 indicates that both of x_1 and x_2 might be related to operators' decision because division times along them, 7 and 4 respectively, are larger than 0. *Fig.2* illustrates the recognition information for operator decision, which is equivalent to $\eta(X \mid D)$ due to our assumption that $p(X)$ is approximately a uniform distribution. It shows that an operator will select a job to be mounted using minimum processing time rule mixed with identical parts size preference. Some domains in Fig.2, whose distribution density equals 0, are invalid because of insufficient data. Therefore generally more past data should be provided if we want to obtain a perfect result.

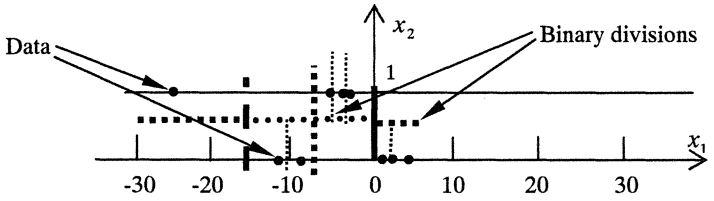

Figure1. Binary divisions performed on 2 dimensions

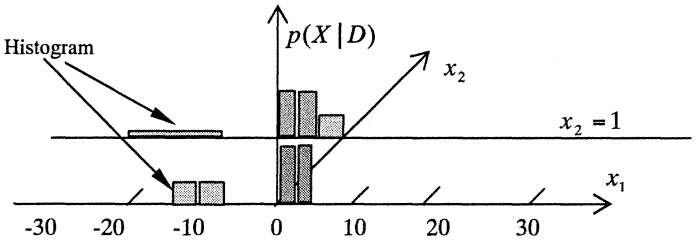

Figure 2. Distribution density after binary division

7. CONCLUSIONS

To recognize operators' decision for a manufacturing system, a model induced from Bayesian thinking is proposed in this paper. We employ non-parametric distribution model to address it and propose a binary division method, whose properties are investigated. It shows that proposed method has the advantage of compressing the model as small as possible, eliminating irrelevant features as well. An example is provided to illustrate the recognition of the operators' decisions.

REFERENCES

1. Anil K. Jain, Robert P. W. Duin, Jianchang Mao, "Statistical Pattern Recognition: A Review", IEEE Trans. On Pattern analysis and machine intelligence, Vol.22, No.1, Jan. 2000, pp4-37.
2. A.G. Gray, A.W. Moore, "'N-Body' Problem in statistical Learning", Advances in Neural Information Processing System 13, May 2001.
3. Avrim L. Blum, Pat Langley, "Select of Relevant Feature and Examples in Machine Learning", Artificial Intelligence, 1997, pp245-271.
4. B. Moghaddam, T. Jebara, A. Pentland, "Bayesian Modeling of Facial Similarity", In Advances in Neural Information Processing System 11, MIT Press, 1999.
5. P.E.Lassila, J.T. Virtamo, "Nearly optimal importance sampling for Monte Carlo simulation of loss system", ACM trans. On modeling and computer simulation, Vol.10, No.4, Oct. 2000, pp326-347.
6. B.W. Silverman, "Density Estimation for Statistic and Data Analysis", Chapman and Hall, London, 1986.
7. G. Terrell, D. Scott, "Variable Kernel Density Estimation", Ann. Statistic, Vol.20, No.3, pp1236-1265, 1992.
8. T.K. Ho, "Nearest Neighbors in Random Subspaces", Lecture Notes in Computer Science: Advances in Pattern Recognition, pp640-648, 1998.