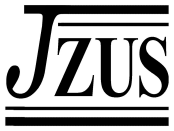
Journal of Zhejiang University SCIENCE
ISSN 1009-3095
http://www.zju.edu.cn/jzus
E-mail: jzus@zju.edu.cn



Quantitative information measurement and application for machine component classification codes*

LI Ling-Feng (李凌丰)†, TAN Jian-rong (谭建荣), LIU Bo (刘 波)

(*State Key Laboratory of CAD&CG, Zhejiang University, Hangzhou 310027, China*)

†E-mail: lilf@mail.hz.zj.cn

Received Feb. 10, 2004; revision accepted Aug. 10, 2004

Abstract: Information embodied in machine component classification codes has internal relation with the probability distribution of the code symbol. This paper presents a model considering codes as information source based on Shannon's information theory. Using information entropy, it preserves the mathematical form and quantitatively measures the information amount of a symbol and a bit in the machine component classification coding system. It also gets the maximum value of information amount and the corresponding coding scheme when the category of symbols is fixed. Samples are given to show how to evaluate the information amount of component codes and how to optimize a coding system.

Key words: Component classification codes, Information source, Information amount, Information entropy of code bit
doi:10.1631/jzus.2005.AS0035 **Document code:** A **CLC number:** TH166

INTRODUCTION

Usually, the following principles are mainly considered for machine component classification coding: (1) Scientificity. The most basic and steadiest attributes or characteristics are selected as the basis of classification codes; (2) Systematization. The attributes or characteristics are sorted systematically and form a rationally scientific classification code system; (3) Expandability. The classifying system is supposed not to fall into disorder when a new component type is added; (4) Compatibility. The system must be consistent with correlated standards, including international standard, national standard, industrial standard, etc.; (5) Synthetic practicality. An overall and rational coding scheme is put forward according to the whole requirement of the manufacturing information system, to meet all the expectations of every department in an organization.

In modern integrated manufacturing environment, a component classification coding system with simple storage, management and exchange of information should be established (Dowlatshahi and Nagaraj, 1998). With the support of computer network and database technology the component classification coding system should have accuracy, consistency and information sharing attributes (Josien and Liao, 2002). These enable the manufacturing system to obtain the most economic and social benefits, and play an indispensable role in the global information and material flow. But, what criterion can be used to estimate the degree of effectiveness and reliability of the machine component classification coding system? And, how to evaluate its quality? Apparently, the key solving these problems is the quantitative measurement of the code information.

The existing references seldom analyze the part classification coding system quantitatively or qualitatively. The process of evaluating and optimizing the coding system often lack definite and quantitative goal. Perotti and Tornincass (1993) considered that classifying is describing parts in multi-dimensional

*Projects supported by the Hi-Tech Research and Development Program (863) of China (No. 2004AA84ts03) and the Science and Technology Committee of Zhejiang Province (No. 2004C31018), China

space depending on various weights. The selected features determine the multi-dimensional vector of a part. To form a conveniently manufacturing part family, they constructed a technology and geometry matrix, and found the most rational grouping status by Ward's algorithm. Ni (1999) established sample fuzzy matrixes, and clustered parts according to different value of λ using maximum tree method, and got different classification instances. Considering the synthesized characters of both structure and processing, Zeng and Li (2000) improved the original coding system by uniting characteristics and increasing adding digits of code. The first author of this paper analyzed the numerical characteristics of classification codes of machine parts quantitatively according to mathematical statistics theory (Li, 2002; Li *et al*., 1999).

With the accelerated promotion of information technology in the manufacturing industry, new and more urgent requests are being put forward for mechanical part classification codes and quantitative analyses of these codes appear (Feng and Li, 2000; Karkkainen *et al*., 2003; Qamhiyah, 1998). Many factors, such as various form, rule, purpose and target of the code, etc. interweave to complicate the situation of mechanical part classification code system. The information theory advanced by C. E. Shannon can be applied to describe information quantitatively with mathematical method (Jiang, 2001). Based on Shannon's Information Theory, the authors regarded codes as information source and put forward a method of quantitatively analyzing part classification codes.

MATHEMEMATICAL MODEL REGARDING CLASSIFICATION CODES AS INFORMATION SOURCE

The codes in part classification code system are generally regarded as a discrete set of limited discrete descriptors (such as characters, letters and numbers) and can be used as discrete information source where these discrete codes appear at random with certain probability. Suppose a_1, a_2, a_3, …, a_r, are possible values of a code bit, e.g. C, and the corresponding probabilities are $p(a_1)$, $p(a_2)$, $p(a_3)$, …, $p(a_r)$. Use discrete random variable C to describe the value of the code bit. Then a mathematical model can be established for the value space vs probability space relationship of a single code bit.

$$[C\cdot P]:\begin{cases} C: & a_1 & a_2 & a_3 & \cdots & a_r \\ P(C): & p(a_1) & p(a_2) & p(a_3) & \cdots & p(a_r) \end{cases}$$

where, $0\le p(a_i)\le 1 \quad (i=1,2,3,...,r)$, $\sum_{i=1}^{r} p(a_i)=1$.

For example, if a classification coding system has 8 code bits denoted by C_k, k=1, 2, ..., s, s=8, respectively, the possible value of each bit is a_i, i=1, 2, …, r, r=10. Specially, each descriptor a_i represents a digit, i.e., a_1=0, a_2=1, a_3=2, a_4=3, a_5=4, a_6=5, a_7=6, a_8=7, a_9=8, a_{10}=9. According to the classification coding rules, these digits have different meaning in different code bit. For example, the classification code of a part is 12008120. According to the coding rules, its features are:

C_1=1 Rings
C_2=2 Nut
C_3=0 Exterior smooth, no steps
C_4=0 No exterior function factors (screw thread, annular groove, etc.)
C_5=8 Other types of exterior end processing
C_6=1 Inside exists as a smooth, single side step through-hole
C_7=2 Interior function factors include screw thread
C_8=0 No aid hole or forming

The statistical probability for a digit in every bit can be found in a batch of parts. The first bit is expressed as follows by information source model.

$$[C_1\cdot P]:\begin{cases} C_1: & 0 & 1 & 2 & \cdots & 9 \\ P(C_1): & 0.14 & 0.486 & 0.32 & \cdots & 0.00 \end{cases}$$

The emergence probability distribution of each code bit is listed in Table 1.

INFORMATION MEASUREMENT OF CLASSIFICATION CODES

Information amount of code value

The information amount of descriptor a_i on bit C_k is described by $I(a_i)$, whose measurement depends on the certainty measurement of a_i. It is known that uncertainty relates to probability, and possibility can be expressed by the value of mathematical expectation. So, $I(a_i)$ must be a function of $p(a_i)$, which is the probability of a_i , viz.

Table 1 The probability distribution of each code bit

$p(a_i)$	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	*Sum*
a_1=0	0.14	0.136	0.448	0.502	0.524	0.192	0.63	0.552	0.3905
a_2=1	0.486	0.226	0.302	0.1	0.054	0.034	0.084	0.036	0.16525
a_3=2	0.32	0.192	0.188	0.23	0.26	0.562	0.252	0.194	0.27475
a_4=3	0.024	0.182	0.016	0.022	0.126	0.048	0	0.078	0.062
a_5=4	0.03	0.128	0.004	0.034	0	0.124	0.022	0.07	0.0515
a_6=5	0	0.054	0.004	0.028	0.024	0.026	0	0	0.017
a_7=6	0	0.008	0.01	0.03	0	0	0.012	0.054	0.01425
a_8=7	0	0.04	0	0.038	0	0	0	0.016	0.01175
a_9=8	0	0.034	0.028	0.016	0.012	0.014	0	0	0.013
a_{10}=9	0	0	0	0	0	0	0	0	0
Average	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1

$$I(a_i)=f[p(a_i)] \qquad (i=1,2,3,...,r)$$

According to Shannon's Information Theory (Jiang, 2001), the function describing information amount is the logarithm of the reciprocal of $p(a_i)$, the emergence probability of a_i, viz.

$$I(a_i)=\log\frac{1}{p(a_i)}=-\log p(a_i) \quad (i=1,2,3,...,r)$$

The unit of $I(a_i)$ depends on the base of the logarithm in the above function. If 2 is the base, the unit will be "bit" (bit–binary unit), viz.

$$I(a_i)=\log_2\frac{1}{p(a_i)} \qquad \text{bit}$$

If 'e' is the base, the unit will be "nat" (nat–natural unit), viz.

$$I(a_i)=\log_e\frac{1}{p(a_i)}=\ln\frac{1}{p(a_i)} \qquad \text{nat}$$

If 10 is the base, the unit will be "Hart" (Hart–Hartley), viz.

$$I(a_i)=\log_{10}\frac{1}{p(a_i)} \qquad \text{Hart}$$

If 'r' is the base, the unit will be 'r' carry information unit, viz.

$$I(a_i)=\log_r\frac{1}{p(a_i)} \qquad r \text{ carry information unit}$$

The formula for converting the logarithmic base of these units can be used to convert them interchangeably. The unit bit will be taken in the following part of this paper unless otherwise stipulated. It means 2 is taken as the logarithm base. For convenience, $\log_2$ is abbreviated to log.

It is easy to find that the probability of the same descriptor a_i varies when a_i emerges on different bit. And the information contained in a_i varies at the same time. Fig.1 shows the relation between $I(a_i)$ and $p(a_i)$.

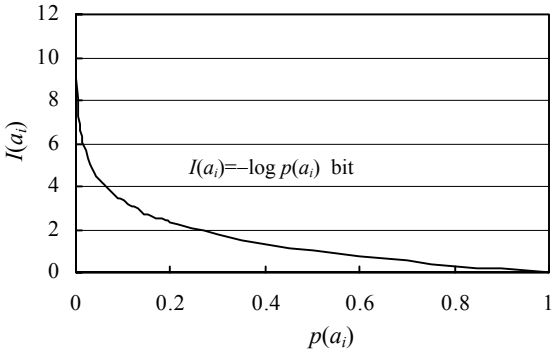


Fig.1 Relation between $I(a_i)$ and $p(a_i)$

Considering two bits C_α and C_β (where α, β=1, 2, …, s; $\alpha\neq\beta$), the probability of C_α taking a_i (i=1, 2, …, r) is $p(a_i)$ and the probability of C_β taking a_j (j=1, 2, …, r) is $p(a_j)$. Assume that the statistical probabilities of C_α and C_β taking a character are independent of one another, then

$$p(a_ib_j)=p(a_i)p(b_j)$$

According to the information amount function,

$$I(a_ib_j)=\log\frac{1}{p(a_ib_j)}=\log\frac{1}{p(a_i)p(b_j)}$$
$$=\log\frac{1}{p(a_i)}+\log\frac{1}{p(b_j)}$$
$$=I(a_i)+I(b_j)\qquad(i,j=1,2,3,...,r)$$

The above formula means that the information amount of two-code-bit descriptor is the sum of their separate information amount. This also holds true in multi code bits condition.

Information amount of code bit

It is possible to apply the information amount function $I(a_i)$ to the measurement of information. But defects still exist. Firstly, the code bit C_k taking the descriptor a_i is a random event with probability $p(a_i)$. The corresponding information amount $I(a_i)$ is also a random variable correlated with the probability value $p(a_i)$. Obviously, it is impractical to measure the information amount of part classification code by using a random variable. Secondly, the function $I(a_i)$ only represents the information amount when C_k takes a certain descriptor a_i. The information amount varies among different descriptors. The information amount function $I(a_i)$ is incapable of measuring the total information of a code bit, determinate variable must be constructed on the base of the information amount function $I(a_i)$.

The variable for measuring the total information of code bits should be an average value of information amount $I(a_i)$ (i=1, 2, …, r) of C_k in its probability space; the probability that C_k taking character a_i (i=1, 2, …, r) is $\{p(a_1), p(a_2), …, p(a_r)\}$. We express the average value by $H(C_k)$. Then

$$H(C_k)=p(a_1)I(a_1)+p(a_2)I(a_2)+...+p(a_r)I(a_r)$$
$$=-p(a_1)\log p(a_1)-p(a_2)\log p(a_2)-...$$
$$-p(a_r)\log p(a_r)$$
$$=-\sum_{i=1}^{r}p(a_i)\log p(a_i)\qquad\text{bit/character}$$

$H(C_k)$ is called the information entropy of code bit C_k and indicates the average information amount supplied by C_k taking a descriptor.

Sample for evaluating the information amount of part codes

From the computing formula for $H(C_k)$ and the probability distribution in Table 1, the information entropy of each code bit can be calculated. For example, assume the probability space of the first code bit is {0.14, 0.486, 0.32, 0.024, 0.03, 0, 0, 0, 0, 0}; then its information entropy $H(C_k)$ can be calculated by using the following formula:

$$H(C_1)=-\sum_{i=1}^{r}p(a_i)\log p(a_i)$$
$$=-0.14\log0.14-0.486\log0.486-$$
$$-0.32\log0.32-0.024\log0.024$$
$$-0.03\log0.03$$
$$=1.709962977\qquad\text{bit/character}$$

Here, it is defined that 0log0=0. And the information entropy of other code bits is shown in Table 2.

Table 2 Information entropy of each code bit

C_k	$H(C_k)$ bit/character
C_1	1.709962977
C_2	2.795182683
C_3	1.864003125
C_4	2.176905956
C_5	1.803495155
C_6	1.897037212
C_7	1.418926983
C_8	1.983303284

The table indicates that the information entropy of each code bit is different and that the information entropy of C_2 is the maximum while that of C_7 is the minimum.

THE OPTIMAL CODING SYSTEM

For a code bit, the r vectors $p_1, p_2, …, p_r$ must satisfy the following constrained equation:

$$\sum_{i=1}^{r}p_i=1$$

The maximum of entropy function $H(p_1, p_2, …, p_r)$ should be a conditional maximum under the constrained conditions.

Construct an assistant function according to the mathematical method of reaching the conditional maximum:

$$F(p_1, p_2, ..., p_r) = H(p_1, p_2, ..., p_r) + \lambda\left[\sum_{i=1}^{r} p_i - 1\right]$$
$$= -\sum_{i=1}^{r} p_i \log p_i + \lambda\left[\sum_{i=1}^{r} p_i - 1\right]$$

where λ is an undetermined coefficient. Finding partial derivatives of function $F(p_1, p_2, \ldots, p_r)$ over the r variables p_i (i=1, 2, …, r) and setting these function values as zero, we get r stable point equations as follows:

$$-(1 + \log p_i) + \lambda = 0 \qquad (i = 1, 2, ..., r)$$

After solving these equations, we get

$$p_i = 2^{(\lambda - 1)} \qquad (i = 1, 2, ..., r)$$

Putting the above formula into the constrained equation, we then get

$$\sum_{i=1}^{r} p_i = \sum_{i=1}^{r} 2^{(\lambda-1)} = r \cdot 2^{(\lambda-1)} = 1$$

So

$$2^{(\lambda-1)} = \frac{1}{r}$$

Now the entropy function $H(p_1, p_2, \ldots, p_r)$ gets the conditional maximum, viz. the corresponding probability distribution of a_i (i=1, 2, …, r).

$$p_i = \frac{1}{r} \qquad (i = 1, 2, ..., r)$$

So the maximum of the entropy function can be obtained as:

$$H_0(p_1, p_2, ..., p_r) = H\left(\frac{1}{r}, \frac{1}{r}, ..., \frac{1}{r}\right) = -\sum_{i=1}^{r} \frac{1}{r} \log \frac{1}{r} = \log r$$

In general, entropy functions of code bits are not larger than the maximum value obtained in the above function, viz.

$$H(p_1, p_2, ..., p_r) \le \log r$$

Thus we get the maximum entropy of code bit. It indicates that in all cases, the code bit takes the maximum information entropy when the probability of any descriptor is equal. And its maximum value is equal to the logarithm of r, the categories of the descriptor that the code bit can take. For maximizing the information entropy, the optimal coding scheme should keep the probability of each descriptor equal.

This function also shows that the maximum value of the code bit's information entropy rests only with the categories of descriptor r. Fig.2 shows the relation between them. The larger r is, the larger the maximum of information entropy is. In the preceding example, code descriptors are selected in the ten digits '0, 1, 2, 3, 4, 5, 6, 7, 8, 9', r=10, so the maximum entropy of the code bit is log10=3.321928095 (bit/character). If there are only two characters 0 and 1, then r=2 and the maximum entropy is log2=1 (bit/character). And maximum entropy becomes log26=4.700439718 (bit/character) while code descriptors are the 26 English letters 'a, b, c, d, …, x, y, z', r=26.

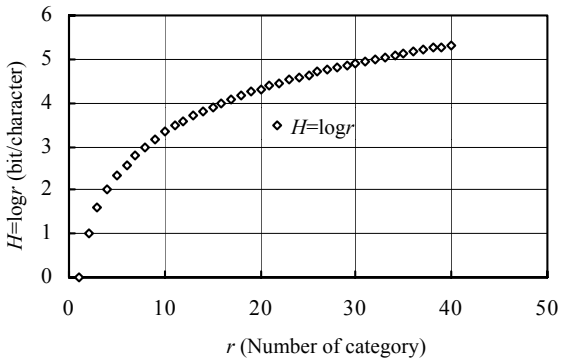


Fig.2 Relation between the maximum information entropy and the number of descriptor categories

CONCLUSION

Because of the uncertainty of part classification code bits in taking its value, any prepared descriptor can possibly emerge on code bits. Such possibility is described by probability. In a batch of part codes, the

information amount reflected by a descriptor has inter-relation with the probability distribution of the descriptor and is a function of probability.

By investigating and analyzing classification codes and regarding them as a form of information, the authors quantitatively measured the information amount using mathematical tool and preserved the mathematical form. In addition to their use in treating and judging the meaning of codes, this study's results have unlimited practical applications by cutting the semantic and pragmatic factors of classification code in the formalization. It is very important that the quantitative evaluating method has nothing to do with the object being coded. It is suitable for any coding system instead of being only adapted to a certain one.

References

Dowlatshahi, S., Nagaraj, M., 1998. Application of group technology for design data management. *Computers and Industrial Engineering*, **34**(1):235-255.

Feng, S.H., Li, J.M., 2000. Component and supplier management technology under the environment of agile enterprise. *Mechanical Science and Technology*, **19**(3): 485-487 (in Chinese).

Jiang, D., 2001. Information Theory and Coding. University of Science and Technology of China Press, Hefei (in Chinese).

Josien, K., Liao, T.W., 2002. Simultaneous grouping of parts and machines with an integrated fuzzy clustering method. *Fuzzy Sets and Systems*, (126):1-21.

Karkkainen, M., Ala-Risku, T., Framling, K., 2003. The product centric approach: a solution to supply network information management problems. *Computers in Industry*, **52**:147-159.

Li, L.F., 2002. Numerical character analysis of the classification codes of machine parts. *Chinese Journal of Mechanical Engineering*, **38**(11):100-104 (in Chinese).

Li, L.F., Tan, J.R, Sha, E.D., 1999. Classification coding of mechanical parts base on feature information. *Mechanical & Electrical Engineering*, **16**(5):190-191 (in Chinese).

Ni, J.F., 1999. Study on clustering of recycling discarded products. *China Mechanical Engineering*, **10**(6):706-708 (in Chinese).

Perotti, G., Tornincass, S., 1993. A semantic technology of part family description and recognition. translated by Li Zhijun. *Group Technology & Production Modernization*, (1): 34-39 (in Chinese).

Qamhiyah, A.Z., 1998. A strategy for the construction of customized design libraries for CAD. *Computer-Aided Design*, **30**(11):897-904.

Zeng, Z.B., Li, R.W., 2000. A method for sorting and classifying machine parts. *Mechanical Science and Technology*, **19**(2):272-274 (in Chinese).