Source program Diagnosis through Pattern Recognition Method

SETSUKO OTSUKI*, KAZUO HATANO** AND YUMIKO KUHARA*

1. *Introduction*

Recently, most computer programs are written in problem oriented languages, such as FORTRAN, ALGOL, etc. In such cases program debugging is not so difficult compared with cases written in assembler languages. Nevertheless computer users are often perplexed by the fact that error messages indicated by language processors do not always point out the correct causes of errors in their source programs because complex subsidiary effects of an error often yield unexpected, too many messages: Which error message indicates what error cause in the source program? Most of users find causes of errors by their experiences, but this is not easy, especially for newcomers.

Then a problem to translate error messages into right causes of errors in the source program is given rise to be considered.

In this paper, we treat this problem from the view point of the pattern recognition using the linear discriminant function on the assumption that the conversational use of a computer is available under its time sharing system. Namely, a diagnosis system is constructed by use of training samples that are obtained by examining "which cause of error yields what error massages" in actual programs presented by many users at the computation center of Kyushu University.

In order to improve the diagnosis rate, a Learning-recognizing system for subsidiary errors is constructed by proposing "clusterability", "confusion parameter" and "region radius" in the information space of which axes denote error messages respectively. In addition, specific problems such as division of error messages, relation between the language processor and the diagnosis system and necessity of conversation which is accompanied to diagnosis are discussed.

2. *Diagnosis of a single error*

When we have an error without subsidiary effects, that is, when only one error message is given for one cause, we regard the cause corresponding to that error message as the most probable source statement error.

This paper first appeared in Japanese in Joho Shori (the Journal of the Information Processing Society of Japan), Vol. 10, No. 4 (1969), 208-215.

* Faculty of Engineering, Kyushu University.
** Faculty of Engineering, Nagoya University.

2.1 *Construction of the diagnosis system*

First, a training sample with a single error message is picked up, then the error message and its corresponding cause are numbered respectively. The j-th error message is denoted by e_j and the i-th cause is denoted by c_i, where $i=1\sim m$ and $j=1\sim n$.

Next, the frequency table of the single error for all pairs of e_j and c_i is prepared. The frequency of the pair e_j and c_i is denoted by F_{ij} as in Fig. 1.

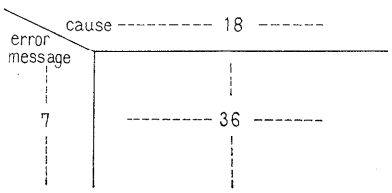

Fig. 1.

2.2 *Diagnosis procedure*

If an error message is given to the diagnosis, the frequency table is searched, and e_j that is equal to the given error message is found. To this e_j, the cause c_i which satisfies

$$F_{ij}=\max_{i'} F_{i'j} \qquad (1)$$

is regarded as the most probable cause.

3. *Diagnosis of complex subsidiary errors*

3.1 *Division of the information space by linear discriminant function*

In the case when complex subsidiary errors are included, that is, when two or more error messages are yielded from one cause, we introduce a method of linear discriminant function.

3.1.1 *Construction of the diagnosis system*

The frequency table is prepared in the similar way as in 2.1 by using training samples which have two or more error messages. As in 2.1, the following notations are introduced.

e_j: the j-th error message in the table.

c_i: the i-th cause in the table.

n: the number of different kinds of error messages.
$(j=1\cdots\cdots n)$

m: the number of different classes of causes.
$(i=1\cdots\cdots m)$

Then, corresponding to a training sample including the i-th cause, the following n-dimensional vector $\boldsymbol{x}_i$ is introduced for all $i=1\sim m$. That is, for $j=1\sim n$

$x_{ji}=1$ If one of error messages of the sample coincides with e_j,

$x_{ji}=0$ otherwise.

$\boldsymbol{x}_i$ means a point belonging to the class i in the n-dimensional information space. The all samples belonging to the class i are given a number ($d=1, 2, \cdots\cdots, d_i$) where d_i is the total number of the samples of the class i. In general, the above number d_i differs in different classes. Through the above procedure, the problem in this section is reduced to a familiar problem to classify new data $\boldsymbol{x}$ through linear discriminant function.

3.1.2 *Diagnosis procedure*

If plural error messages are given, the vector is determined by the method mensioned above. The cause c_i satisfying the following equation is the most probable.

$$[\boldsymbol{x}-1/2(\boldsymbol{\mu}_i+\boldsymbol{\mu}_j)]'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)\geqq 0 \qquad (j=1, 2, \cdots\cdots m) \tag{2}$$

where $\boldsymbol{\mu}_i$ is the mean vector in class i, $\boldsymbol{\Sigma}$ is the equi-variance covariance matrix.

3.2 *Consideration on division of the information space*

Through the linear discriminant function in 3.1, the information space is divided into regions of the same number as the class of the cause. In order to describe the state of each region clearly, the following three quantities are introduced with respect to the d-th sample point $\boldsymbol{x}_{i,d}$ belonging to the i-th cause:

a) Length $G(i/j)$ of the perpendicular drawn to the dividing hyperplane between c_i and c_j,

$$G_d(i/j)\equiv[\boldsymbol{x}_{i,d}-(1/2)(\boldsymbol{\mu}_i+\boldsymbol{\mu}_j)]'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)/\|\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)\| \tag{3}$$

b) Confusion parameter M_i of c_i,

$$M_i\equiv(1/d_i)\sum_{d}^{d_i}\sum_{j\neq i}M_d(i/j)$$

$$\equiv(1/d_i)\sum_{d}^{d_i}\sum_{j\neq i}\{P(i/j)/G_d(i/j)\} \tag{4}$$

where $P(i/j)$ means error probability of c_i to c_j.

c) Clusterability K_{id} of $\boldsymbol{x}_{id}$ to c_i,

$$K_{i,d}\equiv 1/(\boldsymbol{x}_{i,d}-\boldsymbol{\mu}_i)'\Sigma^{-1}(\boldsymbol{x}_{i,d}-\boldsymbol{\mu}_i) \tag{5}$$

Then the state of every region in the information space is classified into the following three:

State-1: Some sample points are scattered outside of the region to which originally they belong (Fig. 2, 2-1). Namely, some points satisfying the following equations exist:

$$G_d(i/j)\leqq 0 \text{ and } M_d(i/j)<0. \tag{6}$$

State-2: All points in a region belong to the same cause and are scattered all over the region (Fig. 2, 2-2). Namely, all points in the region satisfy the following equations:

$$G_d(i/j)>0 \text{ and } M_d(i/j)>0. \tag{7}$$

State-3: All points in a region belong to the same cause and cluster within a

small area in the region (Fig. 2, 2-3). Namely, all points in the region satisfy the following equations:

$$G_d(i/j)>0,\ K_{i,d}\cdot d(i/j)\gg 1 \text{ and } M_d(i/j)\approx 0, \quad (8)$$

where $d(i/j)$ means the Mahalanobis distance.

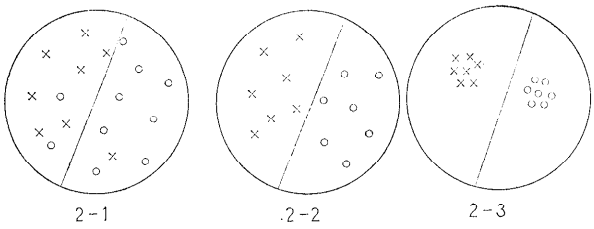

Fig. 2.

In State-1, an erroneous diagnosis occurs essentially. Therefore when a datum is diagnosed to belong to a region originated from c_i, the conversational method, which points out the causes succesively starting from the j-th cause along the direction of the least error probability between c_i and c_j $(j=1\sim m)$ in the region, seems to be the most effective. In this state, even if the right cause is founded or the construction of the learning-recognition system is revised, say, by moving the dividing hyperplane or creating a new region corresponding to new clustering points, the clusterability becomes necessarily smaller because of the increase in the number of such points that correspond to the different causes. As the error massages that induces such a state are usually produced by extremely abridged error processing routine, even an expart in the programming could hardly find the correct cause. The best way to improve the diagnosis rate is to increase the dimension of the information space until the states of all regions change into at least State-2 by increasing the kind of error messages in the language processor.

In State-2, the correct cause can be found by the learning-recognition system. But if new sample points belonging to an unexperienced cause are learnt by the system, the states of the regions may inconveniently transfer to State-1.

In State-3, the system is the best for learning and recognizing, and, moreover, the most of the recently used language processers seem to show this tendency. (See 5.) We introduce the "region radious" R_i by using tne clusterability K_{id} of eq. (5).

$$R_i>\max(1/K_{i,d}),\quad d=1,2,\cdots\cdots d_i \quad (9)$$

$$(\boldsymbol{x}-\boldsymbol{\mu}_i)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i)\leqq R_i. \quad (10)$$

A reduction of the region to the area satisfying both of eqs. (2) and (10) plays the following three important rolls in our learning-recognition system:

1) The remaining hyper subspace between such areas can acept new sample points with unexperienced causes by setting a new region.

2) As will be described in the next section, pseudo information points can be detected in the process of source program diagnosis.
3) The remaining hyper subspace can make a diagnosis of an unknown cause by finding a cluster of the information points.

4. *Source program diagnosis*

The source program diagnosis starts on the assumption that the each error message independently corresponds to a cause as in 2. If there are any massages left that can not be detected or the results of which are not satisfied by users through man-machine conversation, the diagnosis of complex subsidiary errors is continued. In this case, the error messages yielded from a source program generally correspond to plural causes and the correspondence is not clear at all. Therefore all possible combinations between the error mesages must be considered for the diagnosis. That is to say, if r error messages of all coinside with the experienced ones given in the frequency table in 3.1.1, vectors of the number $\sum_{j=2}^{r} {}_rC_j$ are constructed as in 3.1.1. If some of the vectors are not included in the regions satisfying both of eqs. (2) and (10), they are rejected as the pseudo vectors. The causes deduced from the remaining uectors are reported to users successively in the decreasing order of the clusterability by man-machine conversation.

On the contrary, when an error message singly corresponds to an error in the source program, the diagnosis method described in 2 is applied and the causes are reported to users successively in the decreasing order of the frequency. In such a case, as we can see in Table 1 in 5, the more causes correspond to an error, the more times of the man-machine conversation must be repeated. For example, the No. 8 error in Table 1 corresponds to seven kinds of the causes and the probability of detecting the right cause in the first conversation is merely 25 percent and the probabilities of more than the second interaction are all 12.5 percent. In this case it is desirable to return to the design of the language processor and create new error messages corresponding to each cause respectively.

5. *Results of diagnosis*

Table 1. Training samples for a single error.

error No.	cause No.	frequency	error No.	cause No.	frequency
3	22	129	21	36	4
	4	3	25	13	2
	54	1	26	44	3
	34	1		55	1
	57	1	33	43	3
	58	1	27	32	2
	53	1	8	18	2
	98	1		38	1
22	15	15		35	1
	19	4		26	1
4	9	15		7	1
	4	1		39	1
5	2	7		34	1
23	20	7	9	3	1
24	64	6	18	16	1
15	13	7		23	1
	25	1	29	38	1
30	14	6	19	17	1
	27	1	31	24	1
1	31	6	32	43	1
	47	5	12	22	1
	48	2	34	40	1
	46	1	35	59	1
14	11	4	36	21	1
	37	1	11	60	1

Table 2. Training samples with the complex subsidiary errors.

error No. \ cause No.	11		1				55		51	53	50	52	22	12	56	5	14	6	49	99
frequency	4	1	1	1	1	1	3	1	2	3	2	1	1	1	1	1	1	1	1	2
14	1	1																		
1	1									1		1								
3			1	1	1		1	1	1	1	1			1	1				1	
8					1	1						1				1		1		
7							1		1											
6									1						1					
2											1		1						1	
22						1											1	1		
25													1							
10				1																
13					1															
11				1																
26			1																	
4		1																1		1
16								1										1	1	
21														1			1			
17															1					
28																1				
29																		1		
5																				1

Table 3. Results of diagnosis of training samples.

sample No.	error No.					correct cause	result of decision	pass miss	$1/K_{i,d}$	region radius	min $G_d(i/j)$ $j \neq i$	the nearest cause	$\Sigma M_d(i/j)$ $j \neq i$
1	3	2				50	50	P	100.4	150.6	0.015	49	0
2	26	3				1	1	P	139.3	258.5	0.024	55	0
3	11	10	3			1	1	P	121.3	258.5	0.024	55	0
4	13	8	3			1	1	P	172.3	258.5	0.024	55	0
5	22	8				1	1	P	141.0	258.5	0.024	55	0
6	25	2				22	22	P	87.4	131.1	0.122	14	0
7	7	3				51	55	M	80.2	170.4	0.024	1	0
	7	6	3			51	51	P	86.0	129.0	0.262	56	0
8	8	1				52	52	P	123.0	184.5	0.020	53	0
9	3	1				53	53	P	98.6	147.9	0.021	52	0
10	14	4				11	11	P	119.9	179.9	0.095	14	0
11	14	1				11	11	P	69.9	179.9	0.094	14	0
12	7	3				55	55	P	80.2	170.4	0.024	1	0
13	16	3				55	55	P	113.6	170.4	0.024	1	0
14	21	3				12	12	P	115.0	172.5	0.047	52	0
15	17	6	3			56	56	P	69.0	103.5	0.072	50	0
16	28	8				5	5	P	85.0	127.5	0.178	56	0
17	22	21				14	14	P	104.3	156.5	0.061	99	0
18	22	8				6	1	M	141.0	258.5	0.024	22	0
	29	22	16	8	4	6	6	P	79.0	118.5	0.069	14	0
19	16	3				49	55	M	113.6	170.4	0.024	1	0
	3	2				49	50	M	100.4	150.6	0.015	49	0
	16	3	2			49	49	P	92.3	138.5	0.015	50	0
20	4	5				99	99	P	102.8	154.2	0.037	5	0

Table 4. Results of diagnosis of the complex subsidiary, unexperienced errors (ocurrence rate 1.3%).

input error No.			used error No.		correct cause	result of decision	pass miss	$1/K_{i,d}$	region radius	min $G_d(i/j)$ $j \neq i$	the nearest cause	$\Sigma M_d(i/j)$ $j \neq i$
6	16	3	16	3	55	55	P	113.6	170.4	0.024	1	0
7	30	3	7	3	55	55	P	80.2	170.4	0.024	1	0
3	12		3		22	22	P					
2	15	3	2	3	50	50	P	100.4	150.6	0.015	49	0

Table 5. Results of source program diagnosis.

program No.	input error No.	used error No.	correct cause	result of decision	pass miss	remarks
1	15	15	13	13	P	
2	8 25 3	8 25 3	18 13 22	18 13 22	P P P	
3	28 3 30 23 26	3 ⋮ 30 26	60 44	22 ⋮ 14 44	M ⋮ M P	Results of diagnosis with unexperienced cause. (conversation time is 13)
4	22 3	22 3	15 22	15 22	P P	
5	3 1 26	3 1 1 1 1 26	22 44	22 31 47 48 46 44	P M M M M P	
6	5 1	5 1 1	2 47	2 31 47	P M P	
7	7 3 21 25 41	3 ⋮ 25 7 3	55 10	22 ⋮ 13 55	M ⋮ M P	Result of diagnosis with unexperienced cause. (conversation time is 12)
8	26	26	44	44	P	
9	3 21	3	22	22	P	
10	22 1	22 1 1	15 47	15 31 47	P M P	

Table 6. The diagnosis rate of training samples with the complex subsidiary errors (ocurrence rate 9.4%).

The number of times of diagnosis to get the correct cause	frequency	diagnosis rate (%)
1	26	86.7
2	3	10.4
3	1	3.3
4	0	0

Table 7. The diagnosis rate of training samples for a single error (occurrence rate 81.7%).

The number of times of diagnosis to get the correct cause	frequency	diagnosis rate (%)
1	230	87.8
2	18	7.2
3	4	1.5
4	3	1.1
5	2	0.7
6	2	0.7
7	2	0.7
8	1	0.4

Table 8. The diagnosis rate of the complex subsidiary, unexperienced errors (ocurrence rate 1.3%).

The number of times of diagnosis to get the correct cause	frequency	diagnosis rate (%)
1	4	100
2	0	0
irroneous diagnosis	0	0

Table 9. The diagnosis rate of the complex subsidiary errors with unexperienced cause (ocurrence rate 0.9%).

The number of times of diagnosis to decide the unexperienced cause	frequency	diagnosis rate (%)	remarks			
			input error No.			correct cause
3	1	33	28	25	12	7
5	1	33	37	1		63
11	1	33	15	3		62

Table 10. The diagnosis rate of source program diagnosis.

The number of irroneous interaction	frequency	diagnosis rate (%)
0	5	50
1	2	20
4	1	10
12	1	10
13	1	10

6. *Conclusion*

The diagnosis of source programs checked by the error detecting program of a system such as language processors has been discussed. Such a problem would become highly efficient if treated under consideration at the time of designing the error detecting program in a processor.

The experimental results in 5 are obtained by applying to the ALGOL compiler of the computation center of Kyushu University. The detailed explanations are given in the appendix of the original paper in Japanease.