A Short Course at Tamkang University
Taipei, Taiwan, R.O.C., March 7-9, 2006

Classification and clustering methods by probabilistic latent semantic indexing model

Shigeich Hirasawa

Department of Industrial and
Management Systems Engineering
School of Science and Engineering
Waseda University, Japan

hirasawa@hirasa.mgmt.waseda.ac.jp

1. Introduction

Document

Format		Example in paper archives	matrix
Fixed format	Items	- The name of authors - The name of journals - The year of publication - The name of publishers - The name of countries - The year of publication - The citation link	$G \in \{0,1\}^{I \times D}$
Free format	Text	The text of a paper - Introduction - Preliminaries .... - Conclusion	$H \in \{0,1,2,\cdots\}^{T \times D}$

$G = [\, g_{mj} \,]$: An item-document matrix

$H = [\, h_{ij} \,]$: A term-document matrix

d_j : The j-th document

t_i : The i-th term

i_m : The m-th item

g_{mj} : The selected result of the m-th item (i_m) in the j-th document (d_j)

h_{ij} : The frequency of the i-th term (t_i) in the j-th document (d_j)

2. Information Retrieval Model

Text Mining:

- Information Retrieval including
- Clustering
- Classification

Information Retrieval Model

Base	Model
Set theory	(Classical) Boolean Model Fuzzy Extended Boolian Model
Algebraic	(Classical) Vector Space Model (VSM) [BYRN99] Generalized VSM Latent Semantic Indexing (LSI) Model [BYRN99] Neural Network Model
Probabilistic	(Classical) Probabilistic Model Extended Probabilistic Model Probabilistic LSI (PLSI) Model [Hofmann99] Inference Network Model Bayesian Network Model

The Vector Space Model (VSM)

(1)

[Vector Space Model]
Let $\mathcal{T}$ be a term set used for representing a document set $\mathcal{D}$. Let t_i $(i = 1, 2, \cdots, T)$ be the i-th term in $\mathcal{T}$, where $\mathcal{T}$ is a subset of the all term set $\mathcal{T}_0$ appeared in $\mathcal{D}$, and d_j $(j = 1, 2, \cdots, D)$, the j-th document in $\mathcal{D}$. Then a term-document matrix $A = [a_{ij}]$ is given by the weight $w_{ij} \geq 0$ associated with a pair (t_i, d_j). □

Weight w_{ij} is given by

$$w_{ij} = tf(i, j) \cdot idf(i) = a_{ij}$$

$tf(i,j) = f_{ij}$
: The number of the i-th term (t_i) in the j-th document (d_j) (Local weight)

$idf(i,j) = \log(D/df(i))$: General weight

$df(i)$: The number of documents in D for which the term t_i appears

(2)

(term vector) $\boldsymbol{t_i} = (a_{i1}, a_{i2}, \ldots, a_{iD})$: The $\boldsymbol{i}$-th row

(document vector) $\boldsymbol{d_j} = (a_{1j}, a_{2j}, \ldots, a_{Tj})$: The $\boldsymbol{j}$-th column

(query vector) $\boldsymbol{q} = (q_1, q_2, \ldots, q_T)^{\mathrm{T}}$

The similarity $s(q, d_j)$ between $\boldsymbol{q}$ and $\boldsymbol{d_j}$:

$$s(q, d_j) = \frac{\boldsymbol{q}^{\mathrm{T}} \boldsymbol{d}_j}{|\boldsymbol{q}^{\mathrm{T}}| \, |\boldsymbol{d}_j|} \quad \text{(cosine)}$$

The Latent Semantic Indexing (LSI) Model

(1)

[Truncated LSI Model]

Let a term-document matrix $A \in \mathcal{R}^{T*D}$ be given by eq.(2.1). Then the matrix A is decomposed into A_K by the truncated SVD as follows:

$$\begin{aligned} A \to A_K &= \left(U_K \hat{U}\right) \begin{pmatrix} \Sigma_K & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} V_K^{\mathrm{T}} \\ \hat{V} \end{pmatrix} \\ &= U_K \Sigma_K V_K^{\mathrm{T}} \end{aligned}$$

where

$$U_K \in \mathcal{R}^{D*K}$$
$$\Sigma_K \in \mathcal{R}^{K*K}$$
$$V_K \in \mathcal{R}^{T*K}$$

and

$$K \le p \le \max\{T, D\}.$$

In eq.(2.6) $|A - A_K|_F$ is minimized for any K, where p is the rank of A, and $|\cdot|_F$ is the Frobenius matrix norm.

□

(2) Let the term-document matrix A be given by the reduced rank matrix A_K by the truncated SVD, then a query vector $\boldsymbol{q} \in \mathcal{R}^{T*1}$ in eq.(2.4) is represented by $\hat{\boldsymbol{q}} \in \mathcal{R}^{K*1}$ in a space unit dimension K:

(query vector) $\hat{q} = \sum_K^{-1} q \in R^{K \times 1}$

(similality) $s(q, d_j) = \dfrac{\hat{\boldsymbol{q}}^{\mathrm{T}} \hat{\boldsymbol{d}}_j}{|\hat{\boldsymbol{q}}^{\mathrm{T}}| |\hat{\boldsymbol{d}}_j|}$

where

$$\hat{\boldsymbol{d}}_j = \Sigma_K V_K^{\mathrm{T}} \boldsymbol{e}_j \in \mathcal{R}^{K*1}$$

$\boldsymbol{e}_j = (\overset{1}{0}, \overset{2}{0}, \overset{\cdots}{\cdots}, 0, \overset{j}{1}, 0, \cdots, 0)$: the ***j***-th canonical vector

The Probabilistic LSI (PLSI) Model

(1) Preliminary

A) $A=[a_{ij}]$, $a_{ij} = f_{ij}$:the number of a term t_i

B) reduction of dimension similar to LSI

$$K \le \max\{T, D\}$$

C) latent class (state model based on factor analysis)

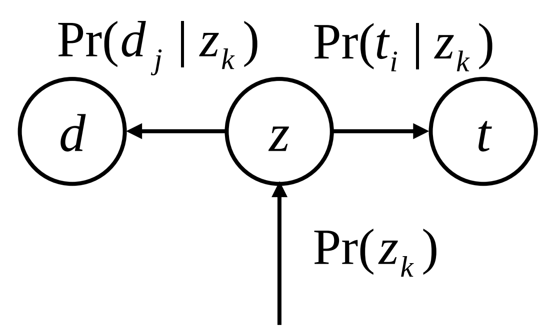
$z_k \in \mathcal{Z}$ $\quad \mathcal{Z}$: a set of states

D) (i) an independence between pairs (t_i, d_j)

(ii) a conditional independence between t_i and d_j

$$\Pr(t_i, d_j) = \sum_{z_k \in \mathcal{Z}} \Pr(d_j) \Pr(t_i|z_k) \Pr(z_k|d_j) \quad (2.10)$$

$$= \sum_{z_k \in \mathcal{Z}} \Pr(z_k) \Pr(t_i|z_k) \Pr(d_j|z_k) \quad (2.11)$$



(2)

[PLSI Model]

Let a term-document matrix $A = [a_{ij}]$ be given by only $tf(i,j)$ of eq.(2.1). Then the probabilities $\Pr(d_j)$, $\Pr(t_i|z_k)$, and $\Pr(z_k|d_j)$ are determined by the likelihood principle, i.e., by maximization of the following log-likelihood function:

$$L = \sum_{i,j} a_{ij} \log \Pr(t_i, d_j) \qquad (2.13)$$

(3) [EM algorithm]

According to eq.(2.11), the maximum value of eq.(2.13) is computed by alternating E-step and M-step until it converges.

E-step:

$$\Pr(z_k|t_i,d_j) = \frac{\Pr(z_k)\Pr(t_i|z_k)\Pr(d_j|z_k)}{\sum_{k'}\Pr(z_{k'})\Pr(t_i|z_{k'})\Pr(d_j|z_{k'})} \qquad (2.14)$$

M-step:

$$\Pr(t_i|z_k) = \frac{\sum_j a_{ij}\Pr(z_k|t_i,d_j)}{\sum_{i',j} a_{i'j}\Pr(z_k|t_{i'},d_j)} \qquad (2.15)$$

$$\Pr(d_j|z_k) = \frac{\sum_i a_{ij}\Pr(z_k|t_i,d_j)}{\sum_{i,j'} a_{ij'}\Pr(z_k|t_i,d_{j'})} \qquad (2.16)$$

$$\Pr(z_k) = \frac{\sum_{i,j} a_{ij}\Pr(z_k|t_i,d_j)}{\sum_{i,j} a_{ij}} \qquad (2.17)$$

Then we have the probabilities $\Pr(d_j), \Pr(t_i|z_k)$, and $\Pr(z_k|d_j)$. □

3. Proposed Method

3.1 Classification method

categories: $C_1, C_2, \ldots, C_K$

(1) Choose a subset of documents $D^* (\subseteq D)$ which are already categorized and compute representative document vectors $\boldsymbol{d}^*_1, \boldsymbol{d}^*_2, \ldots, \boldsymbol{d}^*_K$:

$$\boldsymbol{d}_k^* = \frac{1}{n_k} \sum_{\boldsymbol{d}_j \in C_k} \boldsymbol{d}_j \quad (3.1)$$

where n_k is the number of selected documents to compute the representative document vector from C_k.

(2) Compute the probabilities $\Pr(z_k)$, $\Pr(d_j | z_k)$ and $\Pr(t_i | z_k)$ which maximizes eq.(13) by the TEM algorithm, where $\| Z \| = K$.

(3) Decide the state $z_{\hat{k}} (= C_{\hat{k}})$ for d_j as

$$\max_k \Pr(z_k | d_j) = \Pr(z_{\hat{k}} | d_j) \Rightarrow d_j \in z_{\hat{k}} \quad (3.2)$$

3. Proposed Method

3.2 Clustering method

$||Z|| = K$: The number of latent states

S : The number of clusters

$K \geq S$

(1) Choose a proper K ($\geq S$) and compute the probabilities $\Pr(z_k)$, $\Pr(d_j | z_k)$, and $\Pr(t_i | z_k)$ which maximizes eq.(13) by the TEM algorithm, where $\| Z \| = K$.

(2) Decide the state $z_{\hat{k}} (= c_{\hat{k}})$ for d_j as

$$\max_k \Pr(z_k | d_j) = \Pr(z_{\hat{k}} | d_j) \Rightarrow d_j \in z_{\hat{k}} \qquad (3.3)$$

If S=K, then $d_j \in c_{\hat{k}}$

(3) If $S<K$, then compute a similarity measure $s(z_k, z_{k'})$:

$$s(z_k, z_{k'}) = \frac{\boldsymbol{z}_k^{\mathrm{T}} \boldsymbol{z}_{k'}}{|\boldsymbol{z}_k| |\boldsymbol{z}_{k'}|} \tag{3.4}$$

$$\boldsymbol{z}_k = (\Pr(t_1 | z_k), \Pr(t_2 | z_k), \cdots, \Pr(t_T | z_k))^{\mathrm{T}} \tag{3.5}$$

and use the group average distance method with the similarity measure $s(z_k,z_{k'})$ for agglomeratively clustering the states z_k's until the number of clusters becomes S. Then we have S clusters, and the members of each cluster are those of a cluster of states.

4. Experimental Results

4.1 Document sets

Table 4.1: Document sets

	contents	format	# words T	amount	categorize	# selected document D_L+D_T
(a)	articles of Mainichi news paper in '94 [Sakai99]	Free (texts only)	107,835	101,058 (see Table 4.2)	Yes (9+1 categories)	300 (S=3)
(b)						200〜300 (S=2〜8)
(c)	Question naire (see Table 4.3 in detail)	fixed and free (see Table 4.3)	3,993	135+35	Yes (2 categories)	135+35

4.2 Classification problem: (a)

Conditions of (a)

- Experimental data: Mainichi Newspaper in ‘94 (in Japanese) 300 article, 3 categories (free format only)

Table 4.2: Selected categories of newspaper

category	contents	# articles D_L+D_T	# used for training D_L	# used for test D_T
C_1	business	100	50	50
C_2	local	100	50	50
C_3	sports	100	50	50
total		300	150	150

- LSI : $\boldsymbol{K} = 81$
  PLSI: $\boldsymbol{K} = 10$

Results of (a)

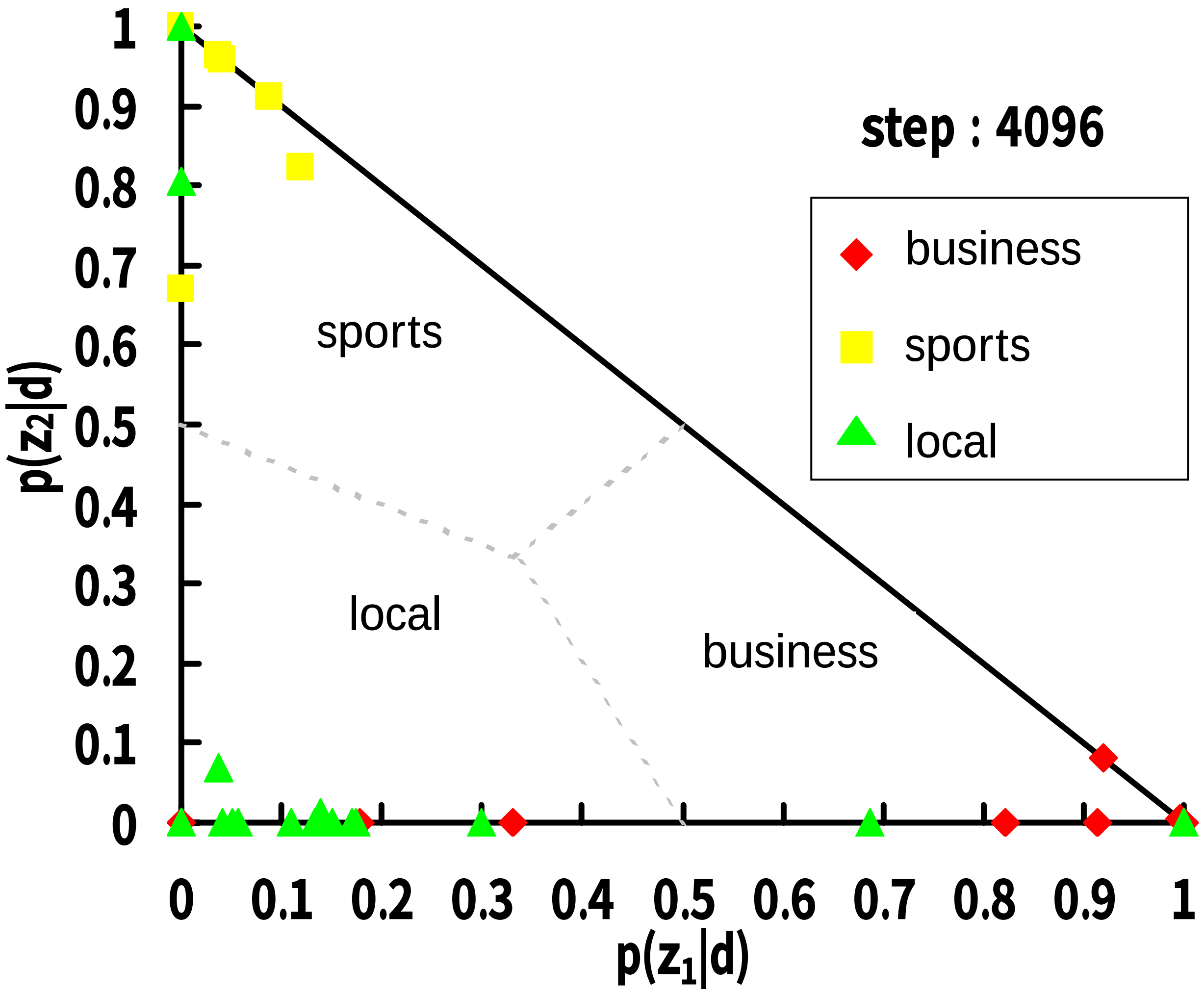
Table 4.5: Classified number form C_k to $C_{\hat{k}}$ for each method

method	from C_k	to $C_{\hat{k}}$		
		C_1	C_2	C_3
VS method	C_1	17	4	29
	C_2	8	38	4
	C_3	15	4	31
LSI method	C_1	16	6	28
	C_2	6	43	1
	C_3	12	5	33
PLSI method	C_1	41	0	9
	C_2	0	47	3
	C_3	13	6	31
Proposed method	C_1	47	0	3
	C_2	0	50	0
	C_3	4	2	44

Table 4.6: Classification error rate

Method	Classification error
VSM	42.7%
LSI	38.7%
PLSI	20.7%
Proposed method	6.0%

Clustering process by EM algorithm



4.3 Classification Problem: (b)

Condition of (b)

We choose S = 2, 3, ..., 8 categories, each of which contains D_L=100〜450 articles randomly chosen. The half of them *DL* is used for training, and the rest of them D_T, for test.

Results of (b)

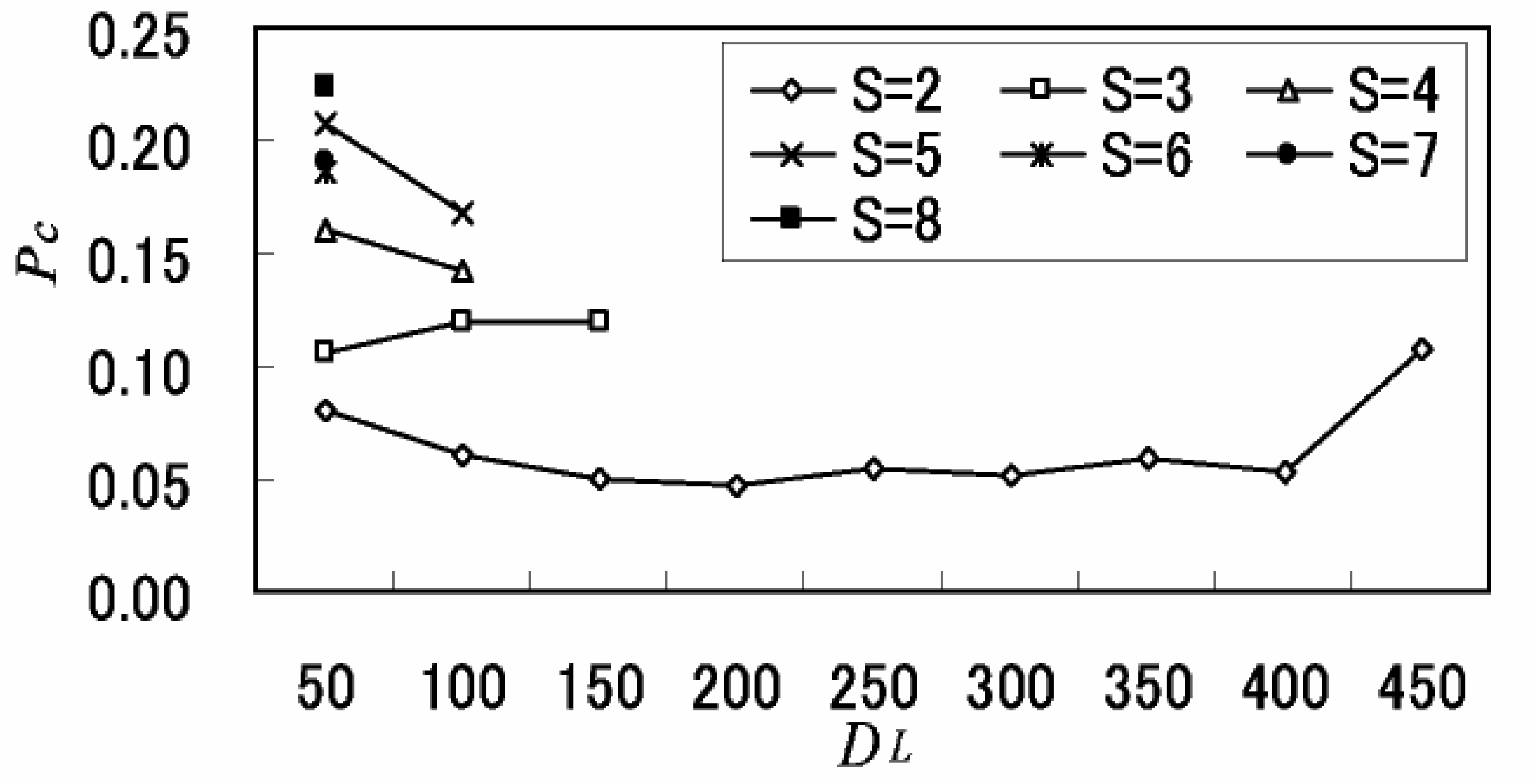


Fig. 4.2: Classification error rate for D_L

Results of (b)

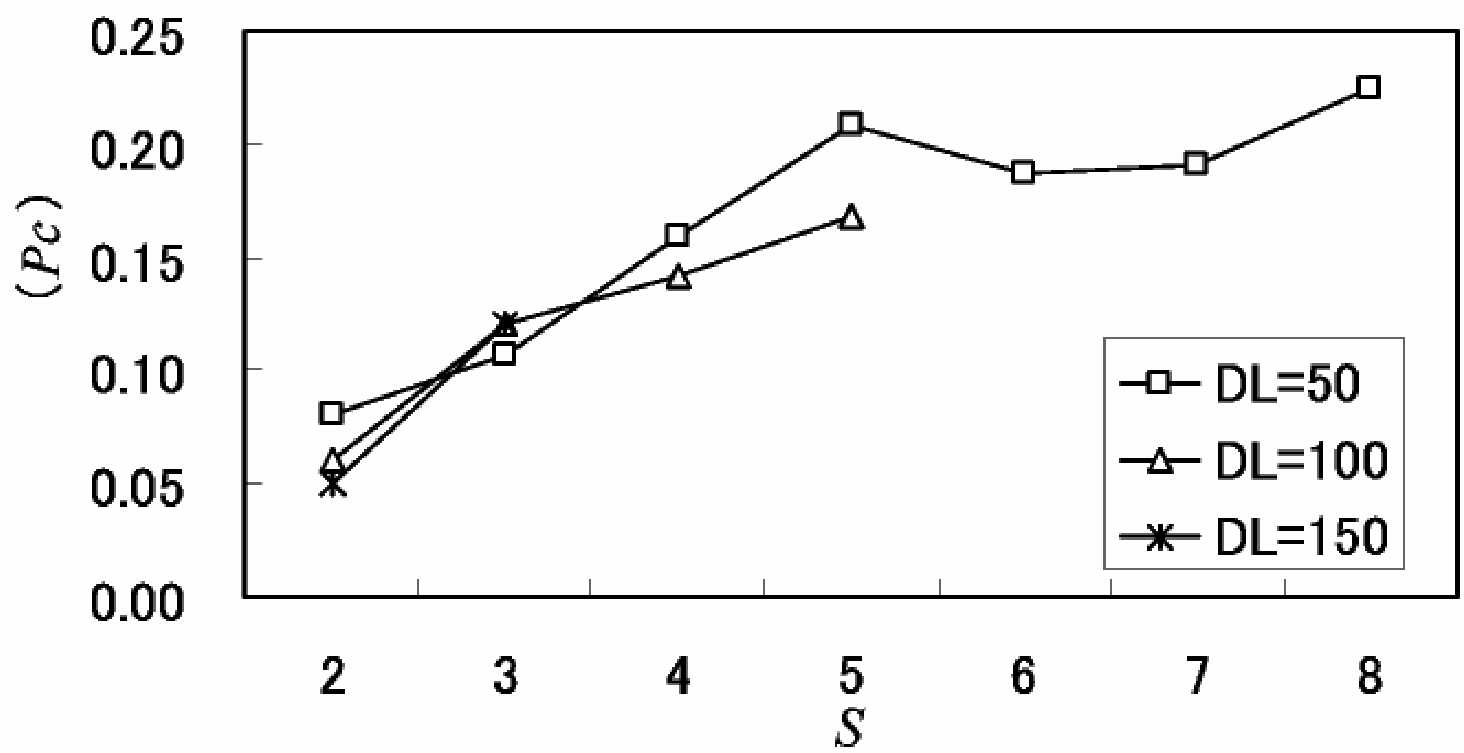


Fig. 4.3: Classification error rate for S

4.4 Clustering Problem: (c)

Student Questionnaire

Table 4.3: Contents of initial questionnaire

Format	Number of questions	Examples
Fixed (item)	7 major questions[2]	- For how many years have you used computers? - Do you have a plan to study abroad? - Can you assemble a PC? - Do you have any license in information technology? - Write 10 terms in information technology which you know[4].
Free (text)	5 questions[3]	- Write about your knowledge and experience on computers. - What kind of job will you have after graduation? - What do you imagine from the name of the subject?

[2] Each question has 4-21 minor questions.
[3] Each text is written within 250-300 Chinese and Japanese characters.
[4] There is a possibility to improve the performance of the proposed method by elimination of these items.

4.4 Clustering Problem: (c)

Object classes

Table 4.4: Object classes

Name of subject	Course	Number of students
Introduction to Computer Science (Class CS)	Science Course	135
Introduction to Information Society (Class IS)	Literary Course	35

Condition of (c)

I) First, the documents of the students in Class CS and those in Class IS are merged.

II) Then, the merged documents are divided into two class (S=2) by the proposed method.

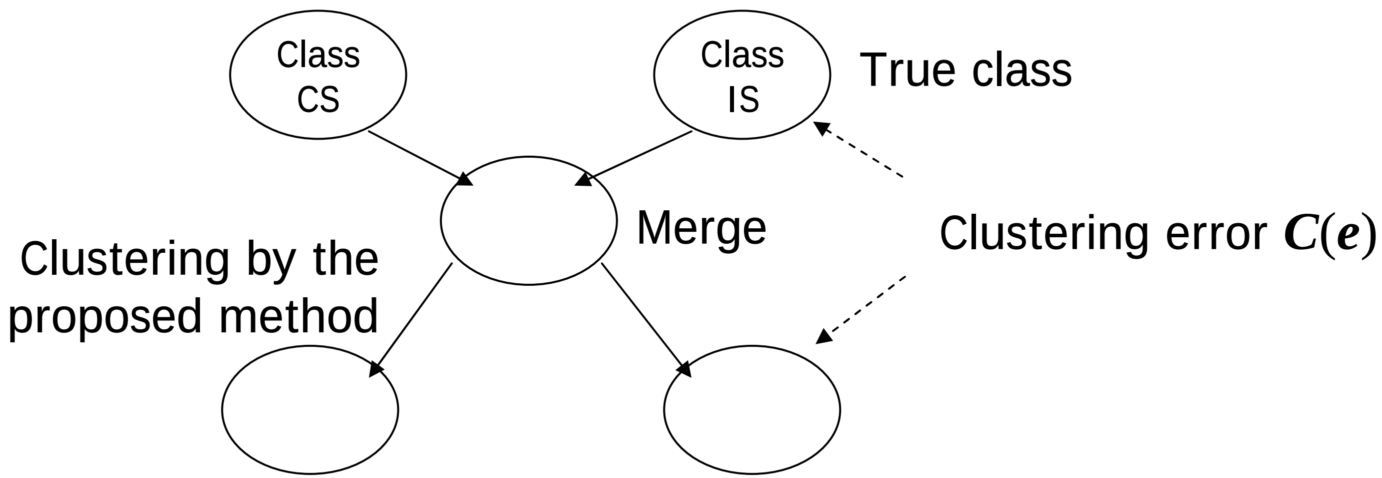


Fig.4.4 Class partition problem by clustering method

Results of (c)

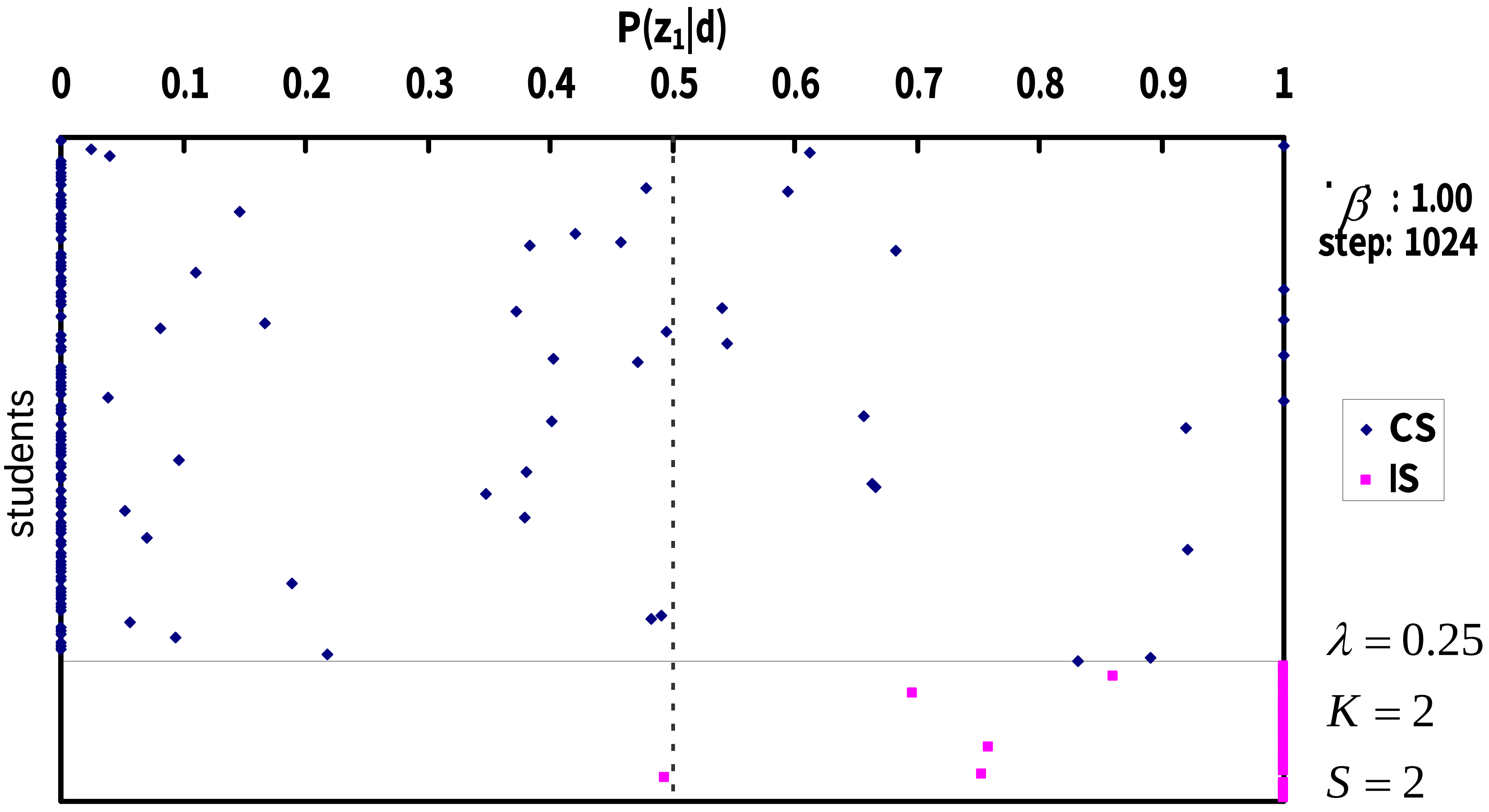
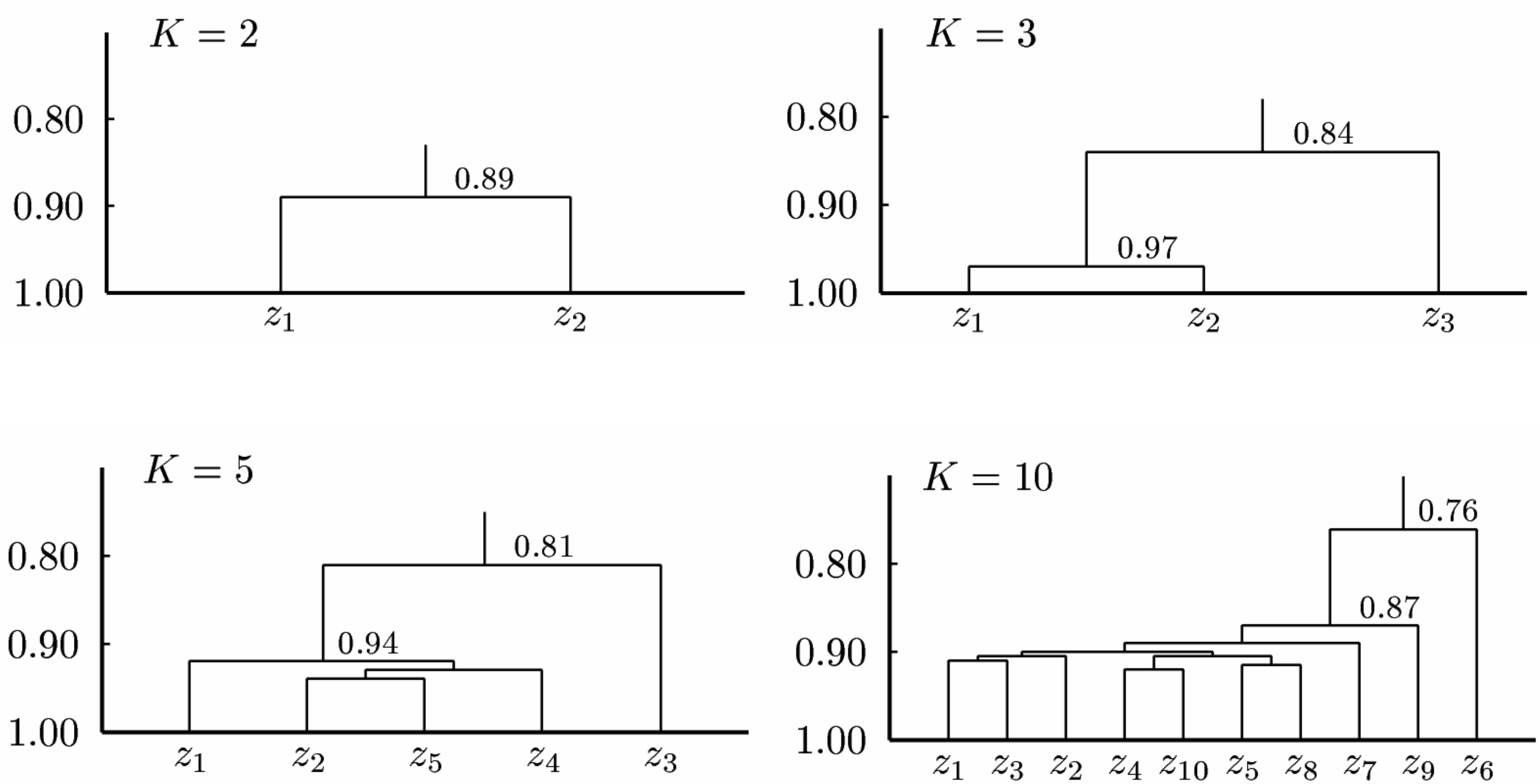
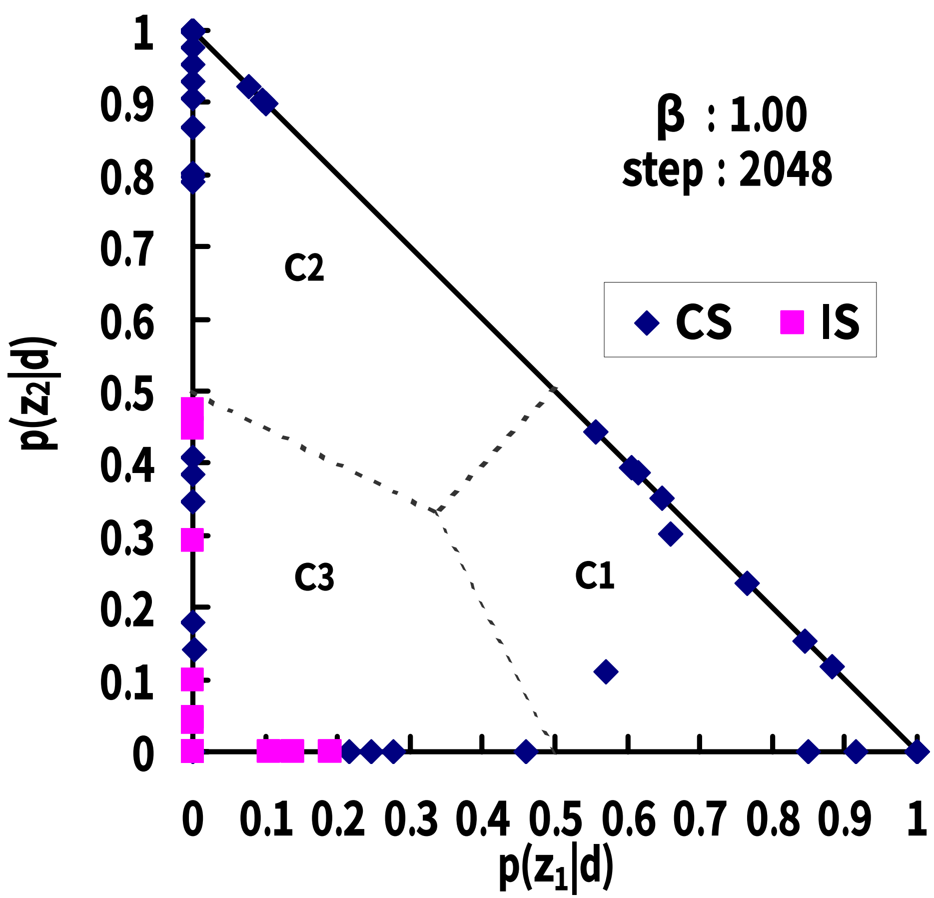


Fig.4.7 Clustering process by EM algorithm, K=2

similarity


K = 2
0.80
0.90
1.00
0.89
z1
z2
K = 3
0.80
0.90
1.00
0.84
0.97
z1
z2
z3
K = 5
0.80
0.90
1.00
0.81
0.94
z1
z2
z5
z4
z3
K = 10
0.80
0.90
1.00
0.76
0.87
z1
z3
z2
z4
z10
z5
z8
z7
z9
z6



$\lambda = 0.25$

$K = 3$

$S = 2$

Fig.4.7 Clustering process for EM algorithm, K=3

K-means method

***S*=*K*=2** **$C(e)$=0.411**

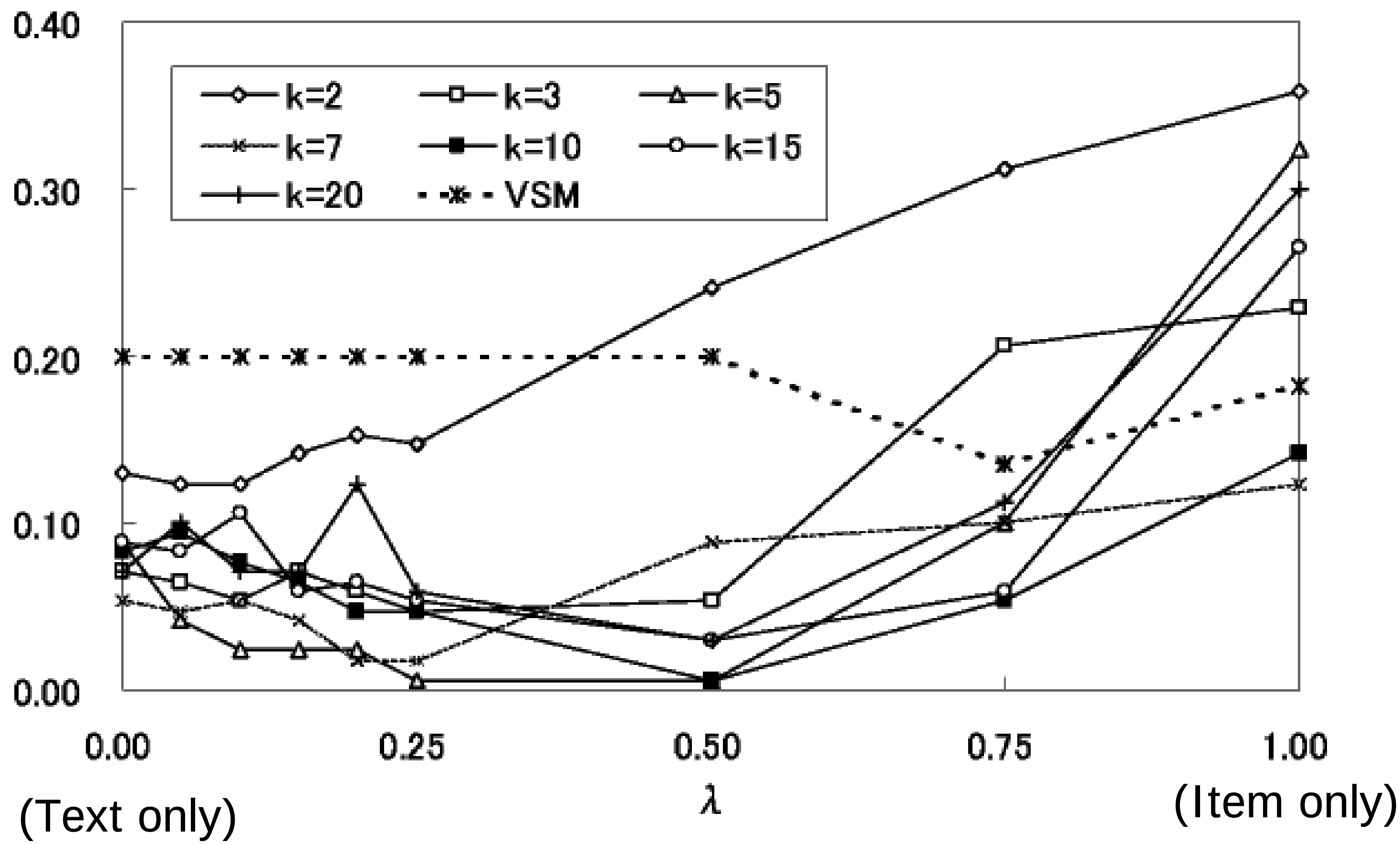


Fig.4.5 Clustering error rate ***C*(*e*)** vs. ***λ***

***C*(*e*)** **:** the ratio of the number of students in the difference set between divided two classes and the original classes to the number of the total students.

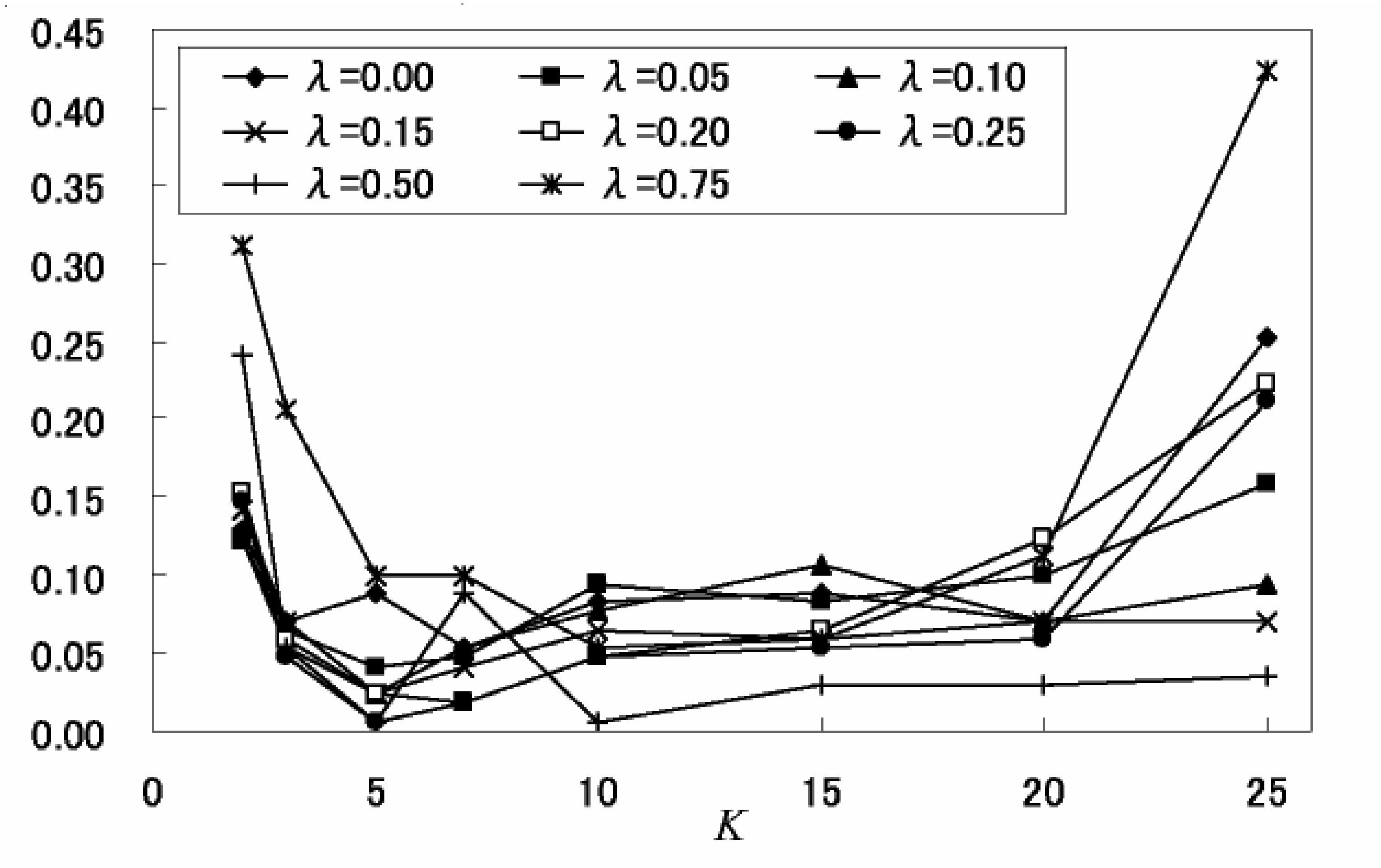


Fig.4.6 Clustering error rate ***C*(*e*)** vs. *K*

Results of (c)

Statistical analysis by discriminant analysis

Table : Characteristics of students for each class by statistical analysis

EV	x_1	x_2	x_3	x_4	x_5
DC	2.411	2.259	1.552	1.336	1.232
Class CS	−	+	+	+	+
Class IS	+	−	−	−	−

EV: Explanatory Variables
DC: Discrimination Coefficient

x_1: This subject is necessary for myself.
x_2: This subject is necessary for the course.
x_3: The main purpose to study is to take for credits.
x_4: I want mid-term test is enforced.
x_5: I want to enter the master course.

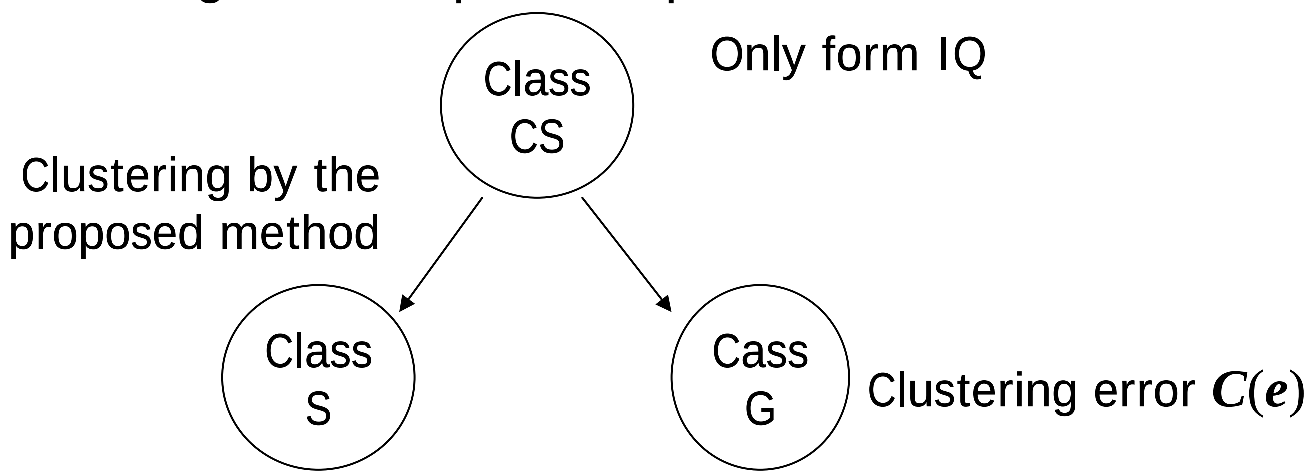
$$z = a_0 + a_1 x_{1j} + a_2 x_{2j} + \cdots + a_5 x_{5j}$$

$$z \geq 0: \quad d_j \in \text{Class CS}$$

$$z < 0: \quad d_j \in \text{Class IS}$$

Another Experiment

Clustering for class partition problem



S: Specialist

G: Generalist

Fig. Another Class partition problem by clustering method

(1) Member of students in each class

	class	Characteristics of students
student's selection	S	- Having a good knowledge of technical terms - Hoping the evaluation by exam
	G	- Having much interest in use of a computer
Clustering	S	- Having much interest in theory - Having higher motivation for a graduate school
	G	- Having much interest in use of a computer - Having a good knowledge of system using the computer

(2) Member of students in each class

Table : Characteristics of students for each class

K	Characteristics of students
2	- No experience in using computers. - High motivation to study the subject.
	- Many experiences in using computer. - Interested in higher grade education and in employment abroad.
3	- Many experiences and knowledge in computer technology. - Low mativation to study the subject
	- High motivation to stydy the subject. - Hihg satisfaction in the class.
5	- High necessity of computers in future. - High level in use of computers in future.
	- Only necessity for credits. - High interest in side job.
10	- High motivation to study the subject. - High scientific sense.
	- Many experiences in using computer.

By discriminant analysis, two classes are evaluated for each partition which are interpreted in table 5. The most convenient case for characteristics of students should be chosen.

5. Concluding Remarks

(1) We have proposed a classification method for a set of documents and extend it to a clustering method.

(2) The classification method exhibits its better performance for a document set with comparatively small size by using the idea of the PLSI model.

(3) The clustering method also has good performance. We show that it is applicable to documents with both fixed and free formats by introducing a weighting parameter λ.

(4) As an important related study, it is necessary to develop a method for abstracting the characteristics of each cluster [HIIGS03][IIGSH03-b].

(5) An extension to a method for a set of documents with comparatively large size also remains as a further investigation.