北海道大学工学部研究報告
第167号（平成6年）

Bulletin of the Faculty of Engineering
Hokkaido University, No. 167 (1994)

Computational Complexity of Subclass Problems

M. Kudo and M. Shimbo

(Received August 31, 1993)

Abstract

When training samples of several classes in an Euclidean space $\mathbf{R}^n$ are given, to find boundaries in $\mathbf{R}^n$ in such a way that they include only samples of a certain class is one of the most important issues in the field of pattern recognition. A subclass problem is one of such problems where boundaries are limited to hyper-rectangles. This paper evaluates the computational complexity of the subclass problem and shows that the problem is NP-hard.

1. Introduction

In the filed of pattern recognition, it is very important for each class to find its discrimination boundaries which can be calculated efficiently in both processes of the construction of boundaries and of judgement of membership for unknown samples, i. e., whether the samples belong to the class or not. Such processes are done on the basis of training samples for given classes. A subclass problem[1] is a kind of such problems : Boundaries are restricted to hyper-rectangles and are required to include only the training samples of a certain class and hold the maximality among such hyper-rectangles. This problem or its similar ones have been appeared several times in the literature of pattern recognition[2,3], and some methods have been proposed to solve those problems optimally or suboptimally. The applicable area includes not only the discrimination of unknown samples but also feature selection[4,5].

In this paper, we discuss the computational complexity of the subclass problem. To do this, we divide the problem into two stages and investigate their complexities separately. One of them is related to a decision problem which is proven to be in the class of NP-complete. Finally, the subclass problem is proven to be NP-hard.

2. Preparation

In this section, we give some notations and definitions. The notations should be referred to reference (6).

2.1. Some classes of problems. A problem Π consists of a set D_Π of *instances*.

A *decision proplem* Π is a problem so as to answer "yes" or "no", given an instance $I \in$

Information Engineering
Email ; mine@huie. hokudai. ac. jp

D_Π. One of well-known decision problems is:

CLIQUE (with K)
INSTANCE: Graph $G=(V, E)$, positive integer $K \leq |V|$.
QUESTION: Does G contain a clique of size K or more, i.e., a subset $V' \subseteq V$ with $|V'| \geq K$ such that every two vertices in V' are connected by an edge of E ?

A *search problem* Π consists of D_Π, and for each instance $I \in D_\Pi$, a set $S_\Pi[I]$ of finite objects called *solutions* for I. An algorithm is said to *solve* a search problem Π if, given as input any instance $I \in D_\Pi$, it returns the answer "no" whenever $S_\Pi[I]$ is empty and otherwise returns some solution $s \in S_\Pi[I]$. A decision problem can be associated to a search problem by answering "yes" if $S_\Pi[I] \neq \emptyset$ and "no" otherwise.

In a search problem Π, each instance $I \in D_\Pi$ has an associated solution set $S_\Pi[I]$, and for the given I, we are required to find *one* element of $S_\Pi[I]$. The *enumeration problem* based on the search problem Π is "Given I, what is the cardinality of $S_\Pi[I]$, i. e., how many solutions are there ?" For example, for CLIQUE, the following is the enumeration problem:

ENUM CLIQUE (with K)
INSTANCE: Graph $G=(V,E)$, positive integer $K \leq |V|$.
QUESTION: How many cliques of size K or more are there for G ?

Furthermore, we introduce a new class of problems. A *list problem* Π is associated with the search problem and requires us to *list* (*display*) all solutions of $S_\Pi[I]$ for a given I.

2.2 Complexity. The enumeration problems associated with NP-complete decision problems are clearly NP-hard, since if we know the answer for an enumeration problem we can easily answer "yes" or "no" for the corresponding NP-complete problem, according to whether the answer is greater than 0 or not. However, some enumeration problems do not belong to the class P, even if the underlying problems belong to P[6]. For this reason, Valiant[7] proposed a new class of complexity called # P-*complete* which contains many enumeration problems associated with NP-complete problems. This class is defined as:

Definition 2.1 (Valiant, 1979). The class # P is of all problems computed by nondeterministic polynomial time Turing machines that have the additional facility of outputting the number of accepting computations. An enumeration problem Π is said to be # P-complete when $\Pi \in$# P and, for all $\Pi' \in$# P, $\Pi' \alpha_T \Pi$, where α_T denotes that there is a *polynomial Turing reduction* from Π' to Π (for the definition of polynomial Turing reduction, see reference (6))

By a similar discussion about the complexity of enumeration problems, we can say that the list problems associated with # P-complete enumeration problems are clearly # P-hard, since if we can list the solutions we can easily answer the number by incrementing a counter by one, instead of listing one solution.

2.3. Polynomial transformability. For proving NP-completeness of a problem Π, it is enough to find an NP-complete problem Π′ and a *polynomial transformation* $f : \Pi' \rightarrow \Pi$, where f can be calculated in polynomial steps in terms of the size of $I \in D_\Pi$ and has obvious correspondence between solutions of Π′ and those of Π. As a similar way, for proving # P -completeness of a problem Π, it is known to be enough to find an # P-complete problem Π′ and a polynomial time *parsimonious transformation* $f : \Pi' \rightarrow \Pi$ which holds the number of solutions[6].

3. Subclass Problem

In this section, we describe SUBCLASS Problem (denoted by **SP** for simplicity). The problem can be written as follows.

SUBCLASS

INSTANCE : $S = (S^+, S^-, d)$, where S^+ and S^- are the collections of d-imensional vectors.

QUESTION : List every subset P of S^+ such that

(1) for all $\mathbf{y} \in S^-$, $\mathbf{y} \notin \mathrm{Rect}(P)$ (*Exclusiveness*), and

(2) for any P' holding exclusiveness, $P \not\subseteq P'$ (*Maximality*),

where Rect(P) denotes the hyper-rectangle which includes every $\mathbf{x} \in P$ within it minimally. We call P a *subclass*.

For example, $S^+ = \{\mathbf{x}_1 = (0,0), \mathbf{x}_2 = (1,3), \mathbf{x}_3 = (3,1)\}$ and $S^- = \{\mathbf{y} = (2,2)\}$ provide an instance $(S^+, S^-, 2)$ of **SP** (See Fig. 1). In this instance, the subsets of S^+ satisfying exclusiveness are six of $\emptyset$, $\{\mathbf{x}_1\}$, $\{\mathbf{x}_2\}$, $\{\mathbf{x}_3\}$, $\{\mathbf{x}_1, \mathbf{x}_2\}$ and $\{\mathbf{x}_1, \mathbf{x}_3\}$, e. g., a rectangle $\mathrm{Rect}(\{\mathbf{x}_1, \mathbf{x}_2\}) = [0, 1] \times [0, 3]$ does not include $\mathbf{y} = (2,2)$. By maximality, the final answer becomes two of $\{\mathbf{x}_1, \mathbf{x}_2\}$ and $\{\mathbf{x}_1, \mathbf{x}_3\}$.

In **SP**, we divide the whole problem into two stages of LIST_EX and LIST_SUB :

(1) LIST_EX ; list exclusive subsets of S^+, and

(2) LIST_SUB ; list maximal elements of the set consisting of all exclusive subsets.

Related to LIST_EX, we consider the following three problems :

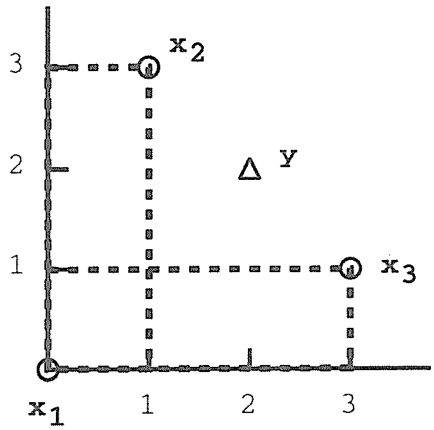

Fig. 1 An instance of **SP**. Dotted lines denote rectangles including samples minimally.

EX (with K)
INSTANCE: $S=(S^+,S^-,d)$, a positive integer $K \leq |S^+|$.
QUESTION: Is there an exclusive subset of size K or more for S?

ENUM_EX (with K)
INSTANCE: $S=(S^+,S^-,d)$, a positive integer $K \leq |S^+|$.
QUESTION: How many exclusive subsets of size K or more are there for S?

LIST_EX (with K)
INSTANCE: $S=(S^+,S^-,d)$, a positive integer $K \leq |S^+|$.
QUESTION: List all exclusive subsets of size K or more for S.

Here, EX is a decision problem, ENUM_EX the corresponding enumeration problem and LIST_EX the corresponding list problem.

Furthermore, we define three problems related to LIST_SUB:

SUB (with K)
INSTANCE: $S=(S^+,S^-,d)$, a positive integer $K \leq |S^+|$.
QUESTION: Is there a subclass of size K or more for S?

ENUM_SUB (with K)
INSTANCE: $S=(S^+,S^-,d)$, a positive integer $K \leq |S^+|$.
QUESTION: How many subclasses of size K or more are there for S?

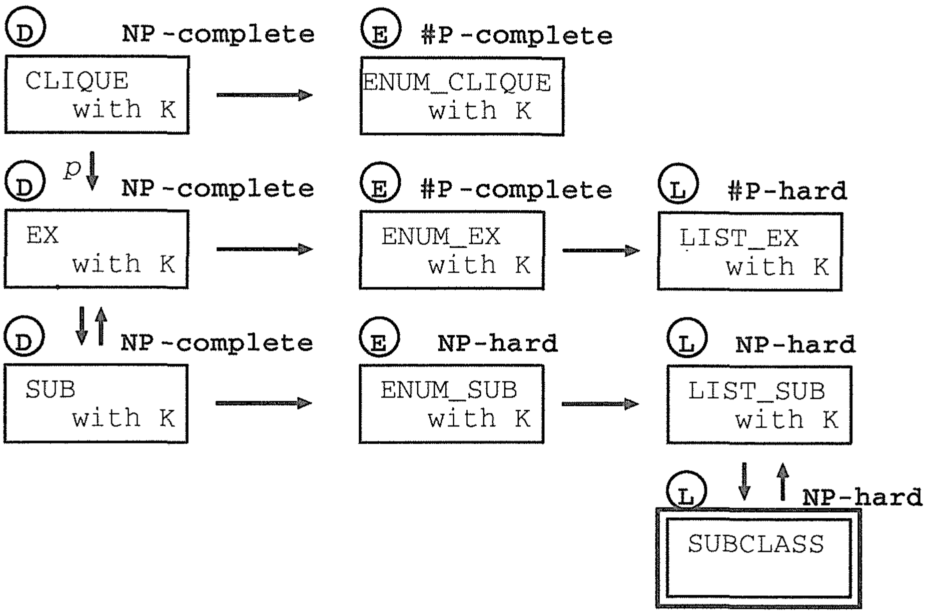

Fig. 2 Diagram of the sequence of transformation used to prove that the subclass problem is NP-hard. Arrows (with label p) denote polynominal (parsimonious) transformations from the sources to the destinations. Characters enclosed by circles denote the classes of the problems, D: decision problems, E: enumeration problems and L: list problems.

LIST_SUB (with K)
INSTANCE: $S=(S^+,S^-,d)$, a positive integer $K \leq |S|$.
QUESTION: List all subclasses of size K or more for S.

Here, SUB is a decision problem, ENUM_SUB the corresponding enumeration problem and LIST_SUB the corresponding list problem.

4. Outline of Proof

Here we give the outline to prove that the subclass is NP-hard. It is done according to Fig. 2.

Since we already know that an enumeration problem is harder to solve than the underlying decision problem and that a list problem is harder to solve than the corresponding enumeration problem, all right-directed arrows in Fig. 2 are proven.

Next, we show that the decision problem SUB is equivalent to EX. This is confirmed easily, because if the answer for the decision problem of SUBCLASS is "yes" ("no") then the answer for EX is also "yes" ("no"), and vice versa. This is because a maximal exclusive subset is a subclass.

We, furthermore, show that LIST_SUB with K is equivalent to SUBCLAS without K.

For any K, if we can list subclasses of size K or more in LIST_SUB, we can easily list all subclasses by setting $K=1$. Inversely, if we can list all subclasses for SUBCLASS, we can list only those of size K or more, by checking their sizes before listing. This check is done in polynomial time in terms of the size of $|S^+|$.

Therefore, the remaining problem to be proven is only that EX is in NP-complete. In addition if the transformation were proven to be parsimonious, then the following fact also could be derived: ENUM_EX is in # P-complete and LIST_EX is in # P-hard. Indeed, this is shown in the following section.

5. Complexity of EX

In this section, we show that the decision problem EX is in NP-complete, that the corresponding enumeration problem ENUM_EX is in # P-complete, and that the corresponding list problem LIST_EX is in # P-hard. For proving this, it is enough to find a parsimonious transformation from the decision problem CLIQUE to EX, because it is known that CLIQUE is in NP-complete already.

Theorem 5.1. EX *is in* NP-*complete and* ENUM_EX *is in* # P-complete.

proof. Let a graph $G=(V,E)$ be an instance of CLIQUE and $V=\{\mathbf{v}_0, \mathbf{v}_2,\ldots, \mathbf{v}_{n-1}\}$. We construct the corresponding instance $S=(S^+,S^-,d)$ of EX as follows. Let $d=n-1$.

1. A set $S^+=\{\mathbf{x}_0, \mathbf{x}_2,\ldots,\mathbf{x}_{n-1}\}$ is constructed as :

$$\mathbf{x}_0=(0,0,\ldots,0)$$

$$\mathbf{x}_2=(1,\frac{1}{2},\ldots,\frac{1}{2})$$

$$\mathbf{x}_2=(\frac{1}{3},1,\frac{1}{3},\ldots,\frac{1}{3})$$

$$\vdots$$

$$\mathbf{x}_i=(\frac{1}{i+1},\frac{1}{i+1},\ldots,\frac{1}{i+1},\overset{\overset{i}{\vee}}{1},\frac{1}{i+1},\ldots,\frac{1}{i+1})$$

$$\vdots$$

$$\mathbf{x}_{n-1}=(\frac{1}{n},\frac{1}{n}\ldots,\frac{1}{n},1).$$

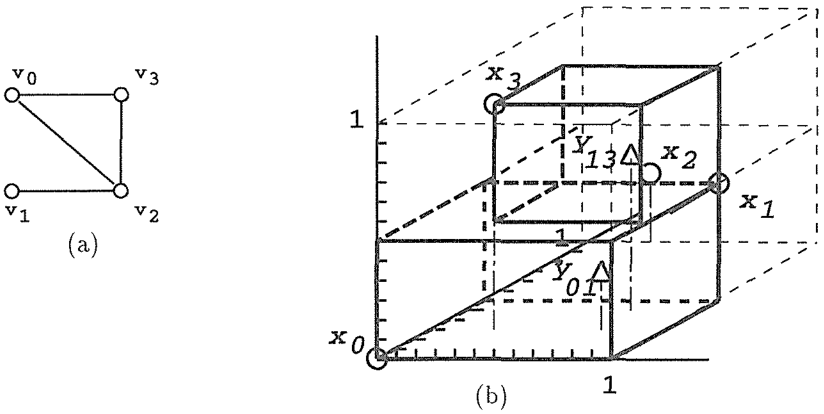

Fig. 3 An example of the correspondence between G and S. (a) $G=(V,E)$, $V=\{\mathbf{v}_0, \mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$ and (b) the corresponding $S=(S^+, S^-, 3)$, where $S^+=\{\mathbf{x}_0=(0,0,0), \mathbf{x}_1=(1,1/2,1/2), \mathbf{x}_2=(1/3,1,1/3), \mathbf{x}_3=(1/4,1/4,1)\}$, $S^-=\{\mathbf{y}_{01}=(3/4,1/8,1/8), \mathbf{y}_{13}=(3/4,3/8,3/4)\}$.

For constructing S^-, we need the following $\mathbf{y}_{ij}$:

$$\mathbf{y}_{0j}=(\frac{1}{2n},\frac{1}{2n},\ldots,\overset{\overset{j}{\vee}}{\frac{3}{4}},\frac{1}{2n},\ldots,\frac{1}{2n})\quad (j\geq 1)$$

$$\mathbf{y}_{ij}=(\frac{1/(i+1)+1/(j+1)}{2},\ldots,\overset{\overset{i}{\vee}}{\frac{3}{4}},\frac{1/(i+1)+1/(j+1)}{2},\ldots,\overset{\overset{j}{\vee}}{\frac{3}{4}},\frac{1/(i+1)+1/(j+1)}{2},$$

$$\ldots,\frac{1/(i+1)+1/(j+1)}{2})\quad (1\leq i<j\leq n-1).$$

Then we construct S^- as follows. S^- is initialized by an empty set. Next, by examining every pair of $(\mathbf{v}_i, \mathbf{v}_j)$, $(i<j)$, we add $\mathbf{y}_{ij}$ into S^- only if $(\mathbf{v}_i, \mathbf{v}_j)\notin E$. An example is shown in Fig. 3.

Under this preparation, we can observe the following relations (see Fig. 3) :

(1) $\mathbf{y}_{ij}\in$ Rect $(\{\mathbf{x}_i,\mathbf{x}_j\})$

(2) $\mathbf{y}_{ij}\notin$ Rect $(\{\mathbf{x}_k,\mathbf{x}_l\})$ for any different $(\mathbf{x}_k,\mathbf{x}_l)$ from $(\mathbf{x}_i,\mathbf{x}_j)$, $(k<l)$

Relation (1) is obvious from the way of construction of $\mathbf{y}_{ij}$. For Relation (2), we use the observation that at least either of the following two relations is always satisfied :

(3) $\text{Rect}_{\bar{i}} \stackrel{\text{def}}{=} \text{Rect}(\{\mathbf{x}_0,\mathbf{x}_1,\ldots,\mathbf{x}_{i-1},\mathbf{x}_{i+1},\ldots\mathbf{x}_{n-1}\}) \supseteq \text{Rect}(\{\mathbf{x}_k,\mathbf{x}_l\})$

(4) $\text{Rect}_{\bar{j}} \stackrel{\text{def}}{=} \text{Rect}(\{\mathbf{x}_0,\mathbf{x}_1,\ldots,\mathbf{x}_{j-1},\mathbf{x}_{j+1},\ldots\mathbf{x}_{n-1}\}) \supseteq \text{Rect}(\{\mathbf{x}_k,\mathbf{x}_l\})$

Here, Relation (3) holds when $\mathbf{x}_k \neq \mathbf{x}_i$ and $\mathbf{x}_l \neq \mathbf{x}_i$, and Relation (4) when $\mathbf{x}_k \neq \mathbf{x}_j$ and $\mathbf{x}_l \neq \mathbf{x}_j$.

Thus, it is sufficient to say that $\mathbf{y}_{ij} \in \text{Rect}_{\bar{i}}$ and $\mathbf{y}_{ij} \in \text{Rect}_{\bar{j}}$. We examine two cases according to $i = 0$ or not. First, we consider the case $i \neq 0$. The interval of $\text{Rect}_{\bar{i}}$ in the i-th dimension is $[0, \max_{\mathbf{x}\in S^+, \mathbf{x}\neq\mathbf{x}_i} x_i]$, where x_i denotes the i-th component of $\mathbf{x}$, and the i-th component of $\mathbf{y}_{ij}$, y_i^{ij}, is 3/4. Since $y_i^{ij} = 3/4 > 1/2 \geq \max_{\mathbf{x}\in S^+,\mathbf{x}\neq x_i} x_i$, we can obtain $\mathbf{y}_{ij} \notin \text{Rect}_{\bar{i}}$. By a similar way, we can confirm $\mathbf{y}_{ij} \notin \text{Rect}_{\bar{j}}$.

Next, let us consider the case $i = 0$. The interval of $\text{Rect}_{\bar{0}}$ in the i-th $(i \neq j)$ dimension is $[1/n,1]$, and $y_i^{0j}=1/2n$. Since $y_i^{0j}=1/2n < 1/n$, we can obtain $\mathbf{y}_{0j} \notin \text{Rect}_{\bar{0}}$. For $\text{Rect}_{\bar{j}}$, by the same reason as the case $i \neq 0$, we can conclude $\mathbf{y}_{ij} \notin \text{Rect}_{\bar{j}}$. After all, Relations (1) and (2) are confirmed.

It should be noted that (2), (3) and (4) indicate that for any subset A of S^+, A includes $\mathbf{x}_i$ and $\mathbf{x}_j$ necessarily if A includes $\mathbf{y}_{ij}$.

Remaining task is to examine the correspondence between solutions for $G=(V,E)$ in CLIQUE and those for $S=(S^+,S^-,d)$ in EX. We consider an exclusive subset $P \in \Phi(S^+,S^-)$, $P = \{\mathbf{x}_{i_1},\mathbf{x}_{i_2},\ldots,\mathbf{x}_{i_t}\}$, $(t \geq 2)$. Let $G_p=(V_p,E_p)$ denote the subgraph of G induced from the vertices corresponding to the elements of P, $V_p=\{\mathbf{v}_{i_1},\mathbf{v}_{i_2},\ldots,\mathbf{v}_{i_t}\}$. Since an exclusive subset whose size is more than 2 is automatically a 2-ary complete subset[8], i. e., any pair of elements of the subset are exclusive, we can conclude that $\text{Rect}(\{\mathbf{x}_i,\mathbf{x}_j,\})$ does not contain any $\mathbf{y} \in S^-$. Thus, every pair $(\mathbf{v}_i,\mathbf{v}_j)$ of V_p belong to E and this means that G_p is a clique.

Inversely, we consider a clique $G'=(V',E')$ of G such as $V'= \{\mathbf{v}_{i_1},\mathbf{v}_{i_2},\ldots,\mathbf{v}_{i_t}\}$ $(t \geq 2)$ and the corresponding subset of S^+, $P_{V'}= \{\mathbf{x}_{i_1}, \mathbf{x}_{i_2},\ldots,\mathbf{x}_{i_t}\}$. Obviously, $P_{V'}$ is 2-ary complete. Is $P_{V'}$ an exclusive subset ? The answer is "yes." We show this as follows. Let us assume $P_{V'}$ is not an exclusive subset. Then, there is $\mathbf{y}_{ij}\in \mathbf{S}^-$ such that $\mathbf{y}_{ij}\in \text{Rect}(P_{V'})$. From the previous argument, this means that $P_{V'}$ must include both of $\mathbf{x}_i$ and $\mathbf{x}_j$ such that $\mathbf{y}_{ij}\in \text{Rect}(\{\mathbf{x}_i,\mathbf{x}_j\})$. This shows that $P_{V'}$ is not 2-ary complete. This is a contradiction.

Now we discuss the case that the size of an exclusive subset or a clique is one. In this case, it is easily confirmed that the exclusive subset is a singleton set and the corresponding clique is an isolated vertex.

By above observation, we could find a parsimonious transformation from G in CLIQUE to S in EX, because it is a one-to-one mapping. Since this transformation can be done in polynomial time in terms of the size of G, the theorem is proven after all. □

6. Conclusion

We treated a problem called a subclass problem and analyzed its computational complexity. As a result, we proved that the problem is NP-hard. Whether the problem is # P-hard or not is still unknown, but it seems to be # P-hard because the subclass problem seems to need in advance to list every exclusive subsets in the worst case and the task is # P-hard as we showed in this paper. These results show that the subclass problem is substantially hard to solve, but this holds for the worst case. The practical problems, especially many problems in pattern recognition, many instances have probably good properties compared to the worst case. Therefore, we have to analyze the complexity of the problem in average or for more limited sets of instances. At present we try to analyze one of the concrete algorithm solving the subclass problem. It will be appeared in one of our further works.

REFERENCES

1) M. Kubo and M. Shimbo, Optimal subclasses with dichotomous variables for feature selection and discrimination. IEEE. *Trans. Systems, Man, and Cybern.*, Vol. 19, pp. 1194-1199, 1989.
2) J. C. Stoffel, A Classifier design technique for discrete pattern recognition problems. IEEE. *Trans. Computer*, Vol. 23, pp. 428-441, 1974.
3) M. Ichino, Pattern classifier design based on the mutual neighborhood graph. Technical report, The faculty of Science and Engineering, Tokyo Denki University, 1984. (in Japanese)
4) M. Kudo and M. Shimbo, Feature selection based on the structural indices of categories. *Pattern Recognition*, Vol. 26, pp. 891-901, 1993.
5) M. Ichino, Optimum feature selection for decision functions. IEICE, Vol. J72-D-II, No. 1, pp. 49-55, 1989. (in Japanese)
6) M. R. Garey and D. S. Johnson, *Computers and Intractability*; A Guide to the Theory of NP-Completeness. W. H. Freeman and Co., San Francisco, 1979.
7) L. G. Valiant, The complexity of computing the permanent. *Theoretical Computer Science*, Vol. 8, pp. 189-201, 1979.
8) M. Kubo, *Analysis of Structures Characterizing Classes and Its Applications*. Doctor thesis, Hokkaido University, April 1988. (in Japanese)