Journal of the Society of Korea Industrial and Systems Engineering
Vol, 28, No. 1, pp.1－7, March 2005.

러프집합이론과 SOM을 이용한 연속형 속성의 이산화

서완석 · 김재련

한양대학교 산업공학과

Discretization of Continuous Attributes based on Rough Set Theory and SOM

Wan-Seok Seo · Jae-Yearn Kim

Industrial Engineering, Hanyang University

Data mining is widely used for turning huge amounts of data into useful information and knowledge in the information industry in recent years. When analyzing data set with continuous values in order to gain knowledge utilizing data mining, we often undergo a process called discretization, which divides the attribute's value into intervals. Such intervals from new values for the attribute allow to reduce the size of the data set. In addition, discretization based on rough set theory has the advantage of being easily applied. In this paper, we suggest a discretization algorithm based on Rough Set theory and SOM(Self-Organizing Map) as a means of extracting valuable information from large data set, which can be employed even in the case where there lacks of professional knowledge for the field.

Keywords : Discretization, Rough Set, SOM

1. 서 론

데이터의 홍수 속에서 유용한 정보를 추출 하기 위해 데이터마이닝(Data Mining)이 다양한 분야에 사용되고 있으며, 대표적인 데이터마이닝 기법으로는 통계적 기법(Statistics), 의사 결정 나무(Decision Tree), 인공신경망(Artificial Neural Network)등이 있다[1][2][3]. 2000년대에 들어서면서 데이터의 모호함과 불확실성에 대한 처리를 위하여 러프집합 이론(Rough Set Theory)을 적용한 다양한 기법들도 소개되고 있으며 데이터마이닝과 같은 지식발견(Knowledge Discovery)에 있어서 효과적인 기법중의 하나로 인정받고 있다[4].

가치 있는 정보로서 가공되기 이전의 원시적 대용량 데이터 집합은 필요 없는 자료들과 불완전하고 부정확한 자료(noise)들을 포함하고 있으며, 동일한 데이터 집합을 같은 기법으로 가공하더라도 데이터 정제와 변형에 따라 도출된 정보의 질이 크게 달라질 수 있기 때문에, 데이터의 전처리과정(Preprocessing)도 지속적인 관심을 끌고 있다. 특히, 현실세계에 존재하는 연속형 속성의 변수들을 이산형 변수로 변형시켜야 할 경우가 발생하기 때문에 이산화 작업에 대한 연구도 활발히 이루어지고 있으며 러프집합 이론을 이용한 이산화 기법들도 소개되고 있다[5][6][7]. 하지만 이산화 과정에서 정보의 손실이 우려되고, 대부분의 처리 대상 데이터들이 대용량이기 때문에 어려움을 겪는 경우가 많다. 가공하고자 하는 데이터 집합에 대해 전문적인 지식이 없거나 축적된 경험이 없는 상태라면 현재 사용되고 있는 이산화 기법들을 적용하기에는 더욱 더 어려움이 있다.

본 연구에서는 정보의 손실을 최소화 하면서 효과적으로 이산화가 가능하도록 러프집합 개념과 SOM(Self-Organizing Map)을 이용하는 알고리즘을 제안하였으며, 실험 결과를 통해 기존 방법과의 차이를 비교 하였다.

2. 러프집합과 SOM

2.1 러프집합

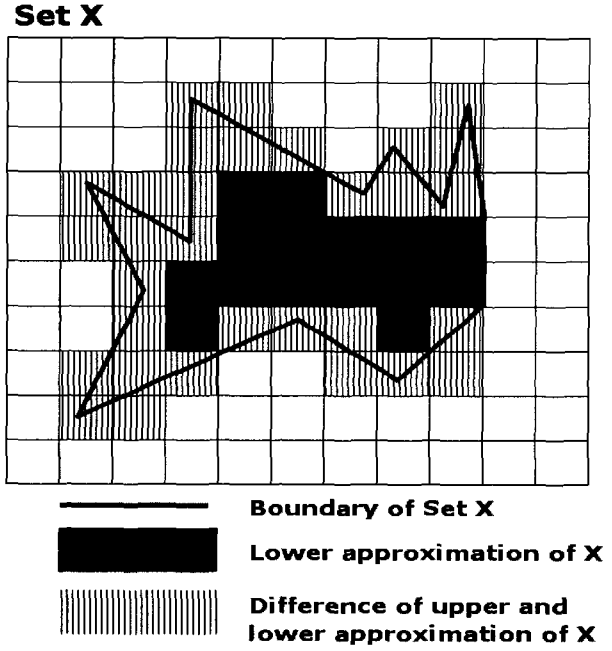
1980년대 초에 Z. Pawlak에 의해 소개된 러프집합 이론은 어떤 집합에서 확실하게 분류되는 하한 근사(Lower Approximation)와 불확실하게 분류되는 상한 근사(Upper Approxima- tion)를 집합 이론을 통해 나타낸다[8]. 하한 근사와 상한 근사에 따라 경계영역(Boundary Region)을 계산 할 수 있으며, 유한집합인 동시에 전체집합인 U와 그 안에서의 동치관계들의 모임 R에 대해 수식으로 나타내면 아래의 식(1), 식(2), 식(3)과 같다. R이 U의 동치관계이면 U/R은 R의 모든 동치류(equivalence class)들의 집합(또는 R에 의한 U의 분류)을 말하고 이를 R의 범주(category of R) 또는 R의 개념(concept of R)이라고 하며 $[x]_R$은 원소 $x \in U$ 를 포함하는 범주를 나타낸다.

$$\text{하한근사 } \underline{R}X = U\{x \in U/R : [x]_R \subseteq X\} \quad (1)$$

$$\text{상한근사 } \overline{R}X = U\{Y \in U/R : [x]_R \cap X \neq \varnothing\} \quad (2)$$

$$\text{경계영역 } BN_R(X) = \overline{R}X - \underline{R}X \quad (3)$$

<그림 1>은 러프집합의 근사 영역에 대한 이해를 돕기 위해 그림으로 나타낸 것이며, 실선으로 표현되어 있는 경계영역에 대해 확실하게 분류되는 하한 근사와 불확실하게 분류되는 상한 근사가 표시 되어 있다.



<그림 1> 러프집합의 근사영역

그리고 이 영역들은 아래의 식(4), 식(5), 식(6)과 같이 정의할 수 있다.

$$X\text{의 } R\text{-긍정영역}(R\text{-positive region}) : POS_R(X) = \underline{R}X \quad (4)$$

$$X\text{의 } R\text{-부정영역}(R\text{-negative region}) : NEG_R(X) = U - \overline{R}X \quad (5)$$

$$X\text{의 } R\text{-경계영역}(R\text{-boundary region}) : BN_R(X) \quad (6)$$

부정영역은 전체집합에서 상한 근사가 제외 된 영역이므로 확실하게 속하지 않는 영역이라고 할 수 있다. 러프집합에서는 경계영역이 적고 하한 근사에 해당되는 범위가 큰 집합일수록 그 집합의 정확도가 높다고 하며, 정확성 척도(accuracy measure)를 이용하여 계산한다. 아래의 식(7)은 전체집합에서 상한근사의 크기 $card\ \overline{R}$ 과 하한근사의 크기 $card\ \underline{R}$ 을 비율로 나타낸 정확성 척도 α 이며, 근사화의 정확도(accuracy of approximation)라고 한다.

$$\alpha_R(X) = \frac{card\,\underline{R}}{card\,\overline{R}} \quad (\text{단}, \ X \neq \varnothing) \quad (7)$$

그리고 같은 조건속성을 갖는 부분집합의 경우, 부분집합의 일관성 정도를 측정할 때는 각 부분집합의 지지도를 이용하여 부분집합의 정확성 척도 β를 계산하며, 아래의 식(8)과 같이 나타낸다. Support의 약자인 supp는 개체들의 빈도이고, r_i는 조건 속성과 결정 속성이 모두 같은 개체이며, r_j는 r_i와 조건부 속성이 같은 개체들이다.

$$\beta = \frac{\text{supp}(r_i)}{\sum \text{supp}(r_j)} \quad (8)$$

또한, 전체집합에 대해서도 얼마나 일관된 정보가 담겨 있는지 알아보기 위해 식(9)와 같이 전체집합의 정확성 척도 γ를 사용하며, 근사화의 질(quality of approximation)이라고 한다.

$$\gamma = \frac{\sum card\ \underline{R}(X_i)}{card\ U} \quad (9)$$

러프집합에 근거한 데이터의 분석과 가공은 행과 열로 구성된 데이터 집합에서부터 출발하며 분류(Classification) 대상이 되는 데이터 집합을 정보 시스템[9]이라 부른다. 일반적으로 수식으로 표현하는 정보시스템 S는 유한한 전체집합 U에 대해 조건 속성 C와 결정 속성 D로 다음

의 식(10)과 같이 표현한다.

$$S=(U,C,D) \tag{10}$$

예를 들어 지구 온난화에 대한 예제[9]로서 아래의 <표 1>을 보면 *Solar Energy*, *Volcanic Activity* 그리고 *Residual CO_2*는 조건 속성이며 *Temperature*는 결정 속성이다. 참고로 *Days Count*는 누적 빈도수를 나타내고 있다.

<표 1> 지구온난화에 대한 정보 시스템

Fact	Solar energy	Volcanic activity	Residual CO_2	Temperature	Days count
1	Medium	High	Low	High	20
2	High	High	High	High	30
3	Medium	Low	High	High	90
4	Low	Low	Low	Low	120
5	High	High	Medium	High	70
6	Medium	Low	High	Low	34

러프 소속 함수[10]는 확률적인 이론을 포함하고 있으며, 개체가 결정 속성의 동치류에 대해 속하는 정도를 확률로 나타낸다. 즉, A라는 정보 시스템에서 집합 X에 대해 $A=(U, A)$이며, $\varnothing \neq X \subseteq U$, $x \in U$일 때, 식(11)과 같이 표현된다. 러프 소속 함수값은 확률에 근거한 값이기 때문에 그 결과는 0에서 1사이의 값을 가진다.

$$\mu_X^A(x)=\frac{|[x]_A \cap X|}{|[x]_A|} \tag{11}$$

여기서 $|\cdot|$는 해당집합의 크기를 의미하며, $\mu_{\varnothing} \equiv 0$이다.

2.2 SOM

튜보 코호넨에 의해 제안된 신경망 모델[11]의 한 분야이며 자기조직화 하는 신경망의 구조가 비교적 단순하다. 역 전파(Back Propagation) 방식의 일반적인 계층적 신경망 모델과는 달리 단순하게 두 개의 층(Layer)으로 이루어져 있으며 입력노드와 출력노드를 연결하는 연결가중치들이 주어진 학습식에 따라 입력패턴에 대하여 스스로 적응해 나가면서 입력벡터들을 그룹핑(grouping)하는 특징이 있다. SOM의 학습식은 아래의 식(13)과 같으며, $x_i(t)$는 시점 t에서의 i번째 입력벡터이고, $w_{ij}(t)$는 시점 t에서의 i번째 입력벡터와 j번째 출력 뉴런 사이의 연결강도이다.

$$d_j=\sum_{i=0}^{N-1}(x_i(t)-w_{ij}(t))^2 \tag{12}$$

또한, i는 0에서 N-1까지의 정수값이다. 가장 유사도가 높은 최소거리 d_j가 구해 졌으면, 선택된 출력노드와의 연결강도(w_{ij})를 재조정하는 과정이 필요하며, 재조정식은 아래의 식(14)와 같다. α는 0과 1사이의 값을 가지는 이득항(gain term)이고, 시간이 경과함에 따라 점차 작아진다.

$$w_{ij}(t+1)=w_{ij}(t)+\alpha(x_i(t)-w_{ij}(t)) \tag{13}$$

SOM을 사용하게 되면 다음과 같은 이점을 얻을 수 있다. 첫 번째로 비지도 학습(unsupervised learning)이기 때문에 특별한 목표치가 필요 없다는 것이다. 예를 들어 군집분석에 사용할 경우, 군집의 수를 정해주지 않아도 스스로 학습하여 군집을 이루게 된다. 두 번째로 학습과정이 1차원이나 2차원의 공간상에서 벡터계산으로 이루어지기 때문에 단순하고 이해하기 쉬우며 계산속도가 매우 빠르다. 세 번째로 결과의 일관성을 들 수 있다. 동일 조건하에서 충분한 계산시간이 주어지면 언제나 동일한 결과의 해를 얻게 된다. 이러한 이점들은 대용량의 데이터를 분석하는 데이터마이닝 기법으로 사용되는데 매우 유용하기 때문에 패턴분류, 군집분석, 경로설정 등의 다양한 분야에 적용되고 있다.

3. 연속형 속성의 이산화

연속형 변수를 이산화 하기 위해 사용되는 전처리과정에서 러프집합 이론과 SOM을 사용하여 이산화 구간을 결정하는 알고리즘을 소개한다. 예제로 사용되는 실제 자동차 데이터베이스[12]는 9개의 조건속성과 1개의 결정속성으로 이루어져 있으며 결정 속성 값은 세 가지의 값 {high, medium, low}으로 나누어져 있다. 자동차의 중량 속성에 대하여 이산화 알고리즘을 적용한다. <표 2>는 예제로 사용할 데이터베이스에서 실제 이산화에 필요한 속성인 자동차의 중량과 마일리지를 추출 한 후, 계산상의 편의를 위해 중량을 기준으로 정렬하여 ID를 부여하였다.

<표 2> 자동차 데이터베이스

ID	weight	mileage
1	698	high
2	786	high
3	798	high
4	867	medium
5	876	medium
6	980	medium
7	980	high
8	987	medium
9	1000	high
10	1002	medium
11	1023	high
12	1039	high
13	1056	medium
14	1094	high
15	1096	medium
16	1098	medium
17	1098	medium
18	1100	medium
19	1123	medium
20	1187	medium
21	1197	medium
22	1557	low
23	1589	medium
24	1600	low

3.1 연속형 속성의 초기 이산화

연속형 속성의 이산화를 위해 우선, 동등 빈도 구간 분할 방법을 이용하여 초기 이산화를 수행한다. 동등 빈도 구간은 근사적으로 같은 개체수를 갖는 구간을 찾아 분할하는 방법으로서, 분할 경계점이 이산화 구간으로 설정 된다. 예제로 사용한 데이터베이스에서는 자동차의 중량이 연속형 변수이기 때문에 이산화 대상 속성으로 선택 되었으며, 자동차의 중량을 동등 빈도 구간으로 이산화 한 결과, 각 구간의 빈도는 3으로 결정 되었고, 첫 번째 구간은 ID 1~3, 두 번째 구간은 ID 4~5와 같이 총 8개의 이산화 구간이 설정 되었다.

3.2 러프 소속 함수값 계산

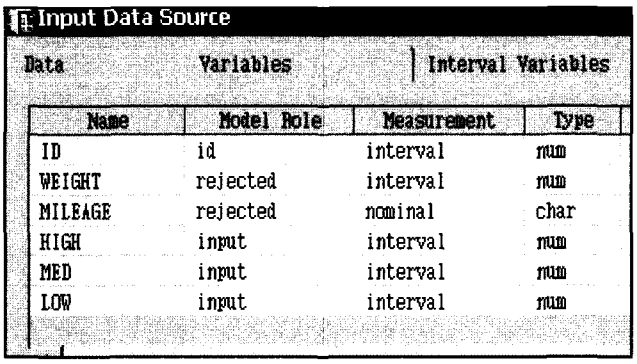
동등 빈도 구간으로 초기 이산화가 완료되었으면, 러프 소속 함수를 이용하여 각 구간에 대해 함수 값을 계산 한다. 결정변수의 값이 세 가지 이상으로 나뉘어져 있다고 하더라도 SOM에서 사용하는 입력변수의 개수에는 제한이 없기 때문에 <표 2>와 같이 결정변수 값의 종류별로 각각 소속 함수값을 계산한다. 예를 들어, ID 7번과 ID 6번의 경우에는 자동차 중량값은 같지만, 초기 이산화 구간에서 ID 6번과 다른 구간에 속해 있기 때문에, 러프 소속 함수 값이 다르게 나왔다.

<표 3 : 러프 소속 함수값을 계산한 결과>

ID	weight	mileage	high	med	low
1	698	high	1	0	0
2	786	high	1	0	0
⋮	⋮	⋮	⋮	⋮	⋮
6	980	medium	0	1	0
7	980	high	0.667	0.333	0
⋮	⋮	⋮	⋮	⋮	⋮
25	1600	low	0	0.333	0.667

3.3 SOM을 이용한 군집 생성

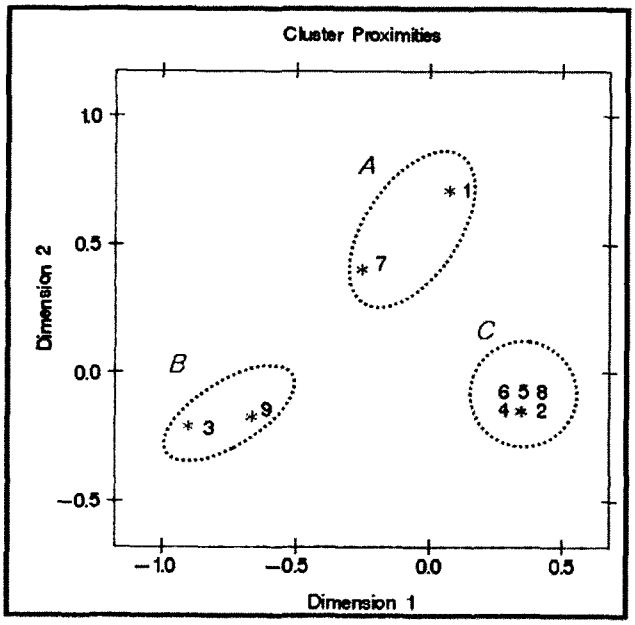
군집화를 통하여 동치류들 간의 유사한 정도를 계산한다. 러프 소속 함수값을 구한 후, 그 결과를 SOM 수행을 위한 입력변수로 사용한다. 예제에서는 아래의 <그림 2>와 같이 ID, 마일리지의 러프 소속 함수값이 SOM의 입력 벡터로 실험이 이루어 졌다.



<그림 2> SOM 입력 변수

3.4 군집 분석을 통한 이산화 구간의 선택

SOM의 수행 결과는 <그림 3>과 같이 2차원 평면상에 나타나게 되므로 군집을 형성하고 각 군집에 따라 이산화구간의 경계값을 결정한다.



<그림 3> SOM 수행 결과

경계 값이 결정될 때에는 중복되는 범위도 발생 할 수 있으나 중복되는 경우에는, 군집내의 빈도에 따라 가중치를 부여하여 범위를 확장하거나 축소하는 방법으로 범위를 결정한다.

실험결과를 먼저 살펴보면, A군집의 최대값은 1094이고 인접한 B 군집의 최소값은 1096이었으므로 첫 번째 이산화 구간의 경계값은 1094와 1096사이가 된다. 그러므로 첫 번째 이산화 구간은 군집내의 최소값 698부터 1094까지가 하나의 구간으로 설정되었다.

3.5 실험결과

실험환경은 Windows 2000 Pro. OS 기반의 Intel Pentium4 2.4GHz CPU에서 SAS E-Miner 4.1을 사용하였다. 실험 대상 데이터베이스는 예제로 소개되었던 자동차 데이터베이스[12]이다. 실험에 사용된 데이터베이스에서 연속형 속성은 자동차의 중량 속성이었으며, 속성 값의 전체범위는 698~1600 이었다. 대상 속성을 이산화 한 후에는 <표 4>와 같이 세 구간으로 나누었으며, 결과적으로 제안한 방법에 의한 이산화의 경우에는 7개의 튜플(tuple)이 감축(reduct)되었다.

<표 4> 이산화 구간 비교

이산화 방법	Chi-Merge 알고리즘	제안한 알고리즘
이산화 대상 속성	weight	weight
이산화 구간	600~800 801~1050 1051~1600	698~1094 1096~1197 1557~1600
지식표현에 필요한 최소 tuple개수	6	5
정확성 척도 γ 값	12%	28%

실험결과를 살펴보면, 총 3개의 군집이 만들어 졌으며 군집 A의 개체들이 가지는 최소값은 698이었고 최대값은 1094이므로, 첫 번째 이산화 구간은 698부터 1094까지 설정되었으며, 군집 B의 최소값은 1096, 최대값은 1197이었으므로 두 번째 이산화 구간은 1096부터 1197까지 이산화 구간으로 설정되었다. 마찬가지 방법으로, 마지막 군집 C의 최대값과 최소값은 각각 1557과 1600이었으므로 이산화 구간은 1557부터 1600까지 하나의 구간으로 설정되었다. 그리고 이산화 이후의 범주형 속성값으로 의사결정표를 간략히 나타내어 데이터베이스의 크기를 감축할 수 있었으며, 그 결과는 <표 5>와 같다.

<표 5> 감축된 의사결정표

weight	mileage	발생횟수
A	High	9
A	Medium	6
B	Medium	7
C	Low	2
C	Medium	1

Chi-Merge 알고리즘을 이용한 결과[13]와 비교해 본다면 데이터 집합의 정확성 척도는 12%에서 28%로 개선되었으며, 정보시스템의 데이터 감축을 비교하면 4%의 개선효과가 있었다. 정확성 척도 γ값은 전체집합에서 동일한 조건 속성값을 가지는 개체들이 같은 결정 속성값을 많이 가질수록 더 높은 값이 나오게 된다.

제안한 알고리즘의 기대효과를 살펴보면, 기존에 제안되었던 Chi-Merge알고리즘의 경우에, 이산화 구간의 설정에 있어서 특정 임계치가 필요하지만, 본 연구에서 제안한 이산화 알고리즘은 이산화 구간에 대한 특정 임계치가 필요 없으므로 해당분야의 전문가가 부족하거나

잘 알려지지 않은 새로운 분야에 대해서도 큰 어려움 없이 적용이 가능하다.

또한 러프집합 이론을 이용하여 초기이산화 결과를 측정해 가면서 이산화 구간을 설정할 수 있기 때문에, 초기이산화 결과에 따라 최종 이산화 구간의 결정에 많은 영향을 미치는 이산화 알고리즘의 단점을 보완할 수 있었다. 그리고 초기이산화 이후에 이산화 구간을 재조정 하는 과정에서는, SOM을 사용하여 빠른 계산속도와 함께 이산화 구간의 개수가 많은 경우에도 어려움 없이 적용이 가능하다는 장점을 얻을 수 있었다.

4. 결 론

본 연구는 대용량 데이터 집합에서 효과적으로 유용한 지식을 추출하는 지식발견 프로세스 단계에 사용되는 전처리 과정중의 하나이다. 일반적으로 분석대상의 데이터베이스에는 불필요한 정보와 반드시 필요한 정보가 함께 포함되어 있으며, 분석기법에 따라 데이터의 속성을 변환시켜야 하는 경우도 흔히 볼 수 있다. 이산화 과정을 거치게 되면 데이터 집합에 따라 지식의 감축도 가능하여 지식발견 기법의 적용에 더 용이할 수 있으며, 전체 데이터 집합의 크기도 함께 감소하는 효과를 얻을 수 있다.

본 연구에서 제안한 알고리즘은 초기이산화를 거치고 러프집합 이론을 이용하여 데이터를 정제한 후, 러프 소속 함수값을 사용하여 SOM의 입력벡터를 산출하고 군집을 생성시켜 각 군집에 대한 분석을 통해 이산화 구간을 설정한다. 알고리즘상의 특정 임계치가 필요 없기 때문에, 데이터 집합에 대한 전문지식이 부족한 경우나 과거에 경험할 수 없었던 분야의 데이터 집합에 대해서도 반드시 임계치를 설정하여 이산화 구간을 결정해야 하는 기존 알고리즘의 단점을 보완하였다. 또한 초기이산화 단계에서 정확성 척도를 이용하여 초기이산화의 결과를 측정 해볼 수 있기 때문에 초기이산화에 큰 영향을 받는 이산화 알고리즘의 단점도 함께 보완할 수 있었다. 그리고 SOM을 사용함으로써 계산속도가 빠르고 입력벡터의 개수에 크게 제한을 받지 않기 때문에 결정속성의 종류가 셋 이상인 경우에도 적용이 용이하다는 장점을 가지고 있다. 그리고 이산화 과정을 수행하는 것만으로도 러프집합의 규칙 추출기법을 사용하면, 추가 작업 없이도 의사결정 규칙의 추출이 가능하다.

기존에 제안되었던 Chi-Merge알고리즘과 수행결과를 비교해 보면 이산화 구간에 대한 특정 임계치가 없어도 이산화가 가능하고, 더 많은 개수의 데이터 감축과 더 우수한 결과의 정확성 척도값을 얻을 수 있었다.

하지만 데이터 집합의 특성에 따라 결정속성과 연관성이 충분하지 않은 경우에 대해서는 다른 이산화 방법들과 마찬가지로 충분한 효과를 얻지 못할 수 있다. 그리고 SOM을 이용한 군집간의 경계설정에 있어서 중복되는 구간의 합리적인 경계 설정에 대한 연구가 더 필요하다.

참고문헌

[1] M. Fayyad, G. Piatesky-Shapiro, P. Smyth, *From Data mining to Knowledge Discovery : An Overview, in Advances in Knowledge Discovery and Data Mining*, pp. 1-34, MIT Press, 1996.

[2] J. Han, M. Kamber, *Data Mining : concepts and Techniques*, pp. 21-26, Morgan Kaufmann publishers, 2000.

[3] M. J. A. Berry , Gordon Linoff, *Data Mining Techniques*, pp. 63-93, John Wiley & Sons, NY, 1997.

[4] S. D. Jitender, V. V. Raghavan, A. Sarkar, H. Sever, "Data Mining : Trends in Research and Development", *Rough Sets and Data Mining analysis of imprecise data*, T.Y.Lin and N.Cercone(Ed.), Kluwer Academic publishers, pp. 9-45, 1997.

[5] J.G. Bazan, H. S. Nguyen, S. H. Nguyen, P. Synak, J. Wroblewski, "Rough set algorithm in classification problem", *Rough set methods and applications : new developments in knowledge discovery in information systems*, L. Polkowski, S. Tsumoto, and T. Y. Lin(Ed.), Physica-Verlag, pp. 57-66, 2000.

[6] W. Ziarko, "Variable Precision Rough Set Model", Journal of Computer & System Sciences, Vol. 46, No 1, pp. 39-59, 1993.

[7] M. J. Beynon, M. J. Peel, "Variable precision rough set theory and data discretisation : an application to corporate failure prediction", Omega, Volume 29, No. 6, pp. 561-576, December 2001.

[8] Z. Pawlak, *Rough sets : Theoretical Aspects of Reasoning About Data*, A Kluwer Academy Publisher, 1991.

[9] Z. Pawlak, "Rough sets and intelligent data analysis", Information Sciences, Volume 147, Issues 1-4, pp. 1-12, November 2002.

[10] Z. Pawlak, A. Skowron, "Rough Membership Functions", *Advances in the Dempster-Shafer theory of evidence*, R. Yager, J. Kacprzyk, and M. Fedrizzi(Ed.). Wiley, pp. 251-271, 1994.

[11] Teuvo Kohonen, *Self-Organizing maps*, 3rd Edition, Springer, 2001.

[12] N. Cercone, H. Hamilton, X. Hu, N. Shan, "Data Mining Using Attribute-Oriented Generation and Information Reduction", *Rough Sets and Data Mining : Analysis of Imprecise Data*, T. Y. Lin and N. Cercone(Ed.), Kluwer Academic publishers, pp. 199-227, 1997.

[13] X. Hu, N. Cercone, "Learning maximal generalized decision rules via discretization : generalization and rough set feature selection", Proceedings ICTAI '97, IEEE, pp. 548-556, 1997.