Journal of the Society of Korea Industrial and Systems Engineering
Vol. 34, No. 2, pp.30－34, June 2011.

구간 데이터를 위한 가변정밀도 러프집합 모형

김 경 택†

한남대학교 산업경영공학과

A Variable Precision Rough Set Model for Interval data

Kyeongtaek Kim†

Dept. of Industrial and Management Engineering, Hannam University

Variable precision rough set models have been successfully applied to problems whose domains are discrete values. However, there are many situations where discrete data is not available. When it comes to the problems with interval values, no variable precision rough set model has been proposed. In this paper, we propose a variable precision rough set model for interval values in which classification errors are allowed in determining if two intervals are same. To build the model, we define equivalence class, upper approximation, lower approximation, and boundary region. Then, we check if each of 11 characteristics on approximation that works in Pawlak's rough set model is valid for the proposed model or not.

Keywords : Variable Precision Rough Set Model, Interval Values, Classification Error

1. 서 론

러프집합이론은 다수의 행과 열로 이루어진 테이블에서 숨겨진 흥미로운 패턴(또는 지식)을 찾아내는데 사용된다. 이러한 러프집합은 다음과 같은 장점을 가지고 있다. 첫째, 부정확 하거나 노이즈(noise)가 있는 데이터로 인하여 일관성이 없게 된 경우에도 적용 가능한 방법론을 제공한다. 러프집합은 근사(approximation)의 개념을 사용하여, 확실한 규칙과 가능한 규칙을 도출할 수 있도록 해주는 장점을 가진다. 둘째, 러프집합은 수학적인 기초위에서 정의되고 활용된다. 따라서, 모호하고 상반된 지식을 다룰 수 있는 러프 집합은 데이터마이닝 분야에서 점차 더 널리 응용되고 있다. 러프집합의 응용분야로는 공학, 의학, 경영, 사회과학 등을 포함한다.

Pawlak[6]이 제시한 러프집합 이론은 여러 분야에 널리 응용되어오고 있음에도 불구하고, 실제적인 응용문제를 해결하는데 있어서 그 한계를 드러냈다[11]. 가장 심각한 문제는 시장조사와 같은 관찰자료를 다룰 때, 타겟 카테고리(예를 들면, 제품 또는 서비스의 구매자들의 카테고리)의 하한근사(lower approximation)가 공집합 아닌 경우가 매우 드물다. 그리고, 대부분의 상한근사(upper approximation)는 전체 집합이 된다[1]. 이러한 한계는 분류(classification) 문제가 원천적으로 비결정적(non-deterministic)이기 때문이다. 이는 이용가능한 정보가 에러가 없는 분류를 허용하지 않음을 의미한다. 예를 들면, 시장조사 결과에 근거하여, 고객의 구매성향을 100% 옳게 예측할 수는 없다.

이러한 문제를 극복하기 위하여, Ziarko[11]는 보다 실제적인 문제에 러프집합모형을 응용 할 수 있도록 기존의 러프집합이론을 일반화시킨 가변정밀도 러프집합 모형(variable precision rough set model)을 제안하였다. 가변정밀도 러프집합 모형에서는 상한근사와 하한근사를

논문접수일 : 2011년 03월 09일　　게재확정일 : 2011년 04월 11일
† 교신저자 kkim610@gmail.com
※ 본 논문은 2010년 한남대학교 학술연구조성비 지원에 의해 연구되었음.

형성하는 기준을 확장시켜, 오분류(classification error)를 허용하였다. 이 때, 허용오분류율(admissible level of classification error rate)을 명시함으로써, 확장의 크기를 제어할 수 있다. 가변 정밀도 러프집합 모형은 Pawlak이 제시한 러프집합 모형의 모든 기본적인 특성을 보존한다[1]. Pawlak의 러프집합 모형은 가변정밀도 러프집합 모형의 특수한 경우로 표현될 수 있으므로, 가변정밀도 러프집합 모형이 러프집합 모형보다 일반적이라 할 수 있다. 최근 들어, 가변 정밀도 러프집합 모형을 응용한 연구가 활발히 진행되고 있다[2, 3, 4, 5, 7, 8, 9, 10].

이제 가변정밀도 러프집합 모형을 속성 값이 구간(interval)을 가지는 객체에 적용하는 경우를 살펴보자. 두 객체의 속성 값이 같은지를 판단할 때 두 구간의 정확한 일치 여부를 사용한다면, 두 객체의 속성 값이 같은 경우는 드물게 발생할 것이다. 이 사실은 객체의 속성 값이 구간을 가지는 경우, 동치관계를 갖는 객체들이 존재하지 않을 가능성이 높아지게 되며, 따라서, 일반집합과 거의 유사한 결과를 가져오게 되어, 러프집합 모형을 사용할 필요가 없게 될 것이다. 따라서, 구간 값이 어느 정도 차이가 나더라도 같은 값으로 보아, 어느 정도의 부정확한 분류를 허용하는 모형의 필요성이 대두된다.

본 논문에서는 두 객체의 속성 값이 구간인 경우 두 객체의 속성 값이 같은지를 판단할 때 오분류를 허용하는 가변정밀도 러프집합모형을 제시한다. 본 논문은 다음과 같이 구성한다. 제 2장에서는 구간 값에 적용 가능한 가변정밀도 러프집합 모형을 제시한다. 이를 위하여, 동치관계를 정의하고, 하한근사 및 상한근사를 정의한다. 제 3장에서는 제시된 모형이 Pawlak이 제시한 러프집합모형의 근사에 관한 기본적인 특성을 보존하는지를 살펴보며, 제 4장에서 연구결과를 요약하고 향후 연구방향을 제시한다.

2. 구간 값을 갖는 가변정밀도 러프집합모형

일반집합 A_1과 A_2를 실수전체집합 U의 부분집합으로 공집합이 아니라 하자. 만일 집합 A_1의 모든 원소가 집합 A_2의 원소이고, 집합 A_2의 모든 원소가 집합 A_1의 원소일 때, A_1과 A_2는 동치관계라고 하며 다음과 같이 표시한다.

$$A_1 = A_2 \qquad \text{if and only if } a \in A_1 \Rightarrow a \subseteq A_2 \text{ and } a \in A_2 \Rightarrow a \subseteq A_1$$

일반적인 집합의 동치관계에 관한 이러한 정의는 너무 엄격하여, 두 집합이 동치 관계는 아니나, 유사한 값을 가지는 경우를 전혀 나타낼 수 없다. 유사한 관계를 표현하기 위해서는 동치관계에 대한 정의를 확장하여야 한다. 먼저, 집합 A_1과 집합 A_2가 동치관계라고 잘못 분류하는 상대적인 크기(relative degree)를 나타내는 척도 $m(A_1, A_2)$를 다음과 같이 정의하자.

$$m(A_1, A_2) = 1 - card(A_1 \cap A_2)/card(A_1 \cup A_2) \quad \text{if} \quad card(A_1 \cup A_2) > 0$$
$$m(A_1, A_2) = 0 \quad \text{if} \quad card(A_1 \cup A_2) = 0$$

여기에서 $card(A_1 \cup A_2)$는 집합 $A_1 \cup A_2$의 원소 갯수를 가리킨다. 이러한 정의는 집합 A_1 및 집합 A_2가 각각 유한집합 일 때에만 성립한다. 구간 값은 실수를 원소로 하는 무한집합이므로, 위의 정의를 사용할 수 없다. 따라서 집합 A_1 및 집합 A_2가 각각 구간일 경우 집합 A_1와 집합 A_2가 동일하다고 잘못 분류하는 상대적인 크기를 나타내는 척도 $m(A_1, A_2)$을 다음같이 정의하자.

$$m(A_1, A_2) = 1 - length(A_1 \cap A_2)/length(A_1 \cup A_2) \quad \text{if} \quad length(A_1 \cup A_2) > 0$$
$$m(A_1, A_2) = 0 \text{ if } \quad length(A_1 \cup A_2) = 0$$

여기에서 $length(A_1)$는 집합 A_1의 구간의 길이를 가리킨다. 다시 말하면, 만일 집합 A_1이 집합 A_2와 동일하다고 한다면, 집합 $A_1 \cup A_2$의 원소 중 $m(A_1, A_2) \times 100\%$의 경우에 대하여 동일한 값이 상대편 집합에 존재하지 않는 오류가 발생하게 된다. 따라서, 척도 $m(A_1, A_2)$는 동치관계의 상대적 오류를 나타낸다.

상대적 오분류의 척도를 사용하여, 집합 A_1과 집합 A_2가 동치인 관계를 다음과 같이 나타낼 수 있다.

$$A_1 = A_2 \text{ if and only if } m(A_1, A_2) = 0$$

오분류 값을 이용하여 유사한 두 구간 A_1와 A_2가 동치관계에 있다는 사실을 표현해보자. 물론 이 경우, 오분류를 허용하고 있으나, 오분류 값 $m(A_1, A_2)$이 너무 크면, 실제적인 문제에의 적용에 의미가 반감될 것이다. $m(A_1, A_2)$가 작을수록 보다 정밀한 동치관계가 이루어지고 있음을 나타낸다.

구간 오분류 허용치가 γ인 동치관계(구간 오분류 허용 동치관계라 부르자) $=_\gamma$를 다음과 같이 정의하자.

$$A_1 =_\gamma A_2 \text{ if and only if } m(A_1, A_2) \leq \gamma$$

그러면, 구간 오분류 허용치 γ가 0인 동치관계 $=_{\gamma=0}$는 원래의 동치관계 = 를 나타낸다. 구간 오분류 허용 동치관계 $=_\gamma$는 다음의 성질을 갖는다.

1) 구간 오분류 허용 동치관계는 반사성(reflexive)을 갖는다. 즉 $A_1 =_\gamma A_1$이 항상 성립한다.
2) 구간 오분류 허용 동치관계는 대칭성(symmetric)을 갖는다. 즉 $A_1 =_\gamma A_2$이면 항상 $A_2 =_\gamma A_1$이 성립한다.
3) $\gamma > 0$인 구간 오분류 허용 동치관계는 이행성(transitive)을 갖지 않는다. 즉 $A_1 =_\gamma A_2$이고 $A_2 =_\gamma A_3$이면 $A_1 =_\gamma A_3$가 성립하지 않는다.
4) 만일 $\gamma_1 < \gamma_2$이고 $A_1 =_{\gamma_1} A_2$이면, $A_1 =_{\gamma_2} A_2$이다.

A = (U, A)가 정보시스템이며, $W \subseteq U$이고 $B \subseteq A$이라 하자. 동치관계의 구간 오분류 허용치가 γ일 때, 원소 $x \in U$의 동치류(equivalence class) $[x]_B$를 다음과 같이 정의한다.

$$[x]_B = \{u \in U \mid \forall a \in B \;\; a(x) =_r a(u)\}$$

또는

$$[x]_B = \{u \in U \mid \forall a \in B \;\; m(a(x),\, a(u)) \le \gamma\}$$

위 정의에 따르면, 동치류는 구간 오분류 허용 동치관계에 있는 구간들의 집합이다.

이제, 하한근사와 상한근사를 정의하자. 동치관계에 대한 구간 오분류 허용치가 γ일 때 W의 B 하한근사 $\underline{B}_\gamma(W)$는 다음과 같이 정의한다.

$$\underline{B}_\gamma(W) = \{x \mid \forall y \in [x]_B \;(\exists w \in W (\forall b \in B \;(b(y) =_\gamma b(w))))\}$$

어떤 원소의 동치류에 속한 모든 원소 각각에 대하여, W에 있는 원소와 B의 관점에서 구간 오분류 허용 동치관계에 있는 원소가 적어도 하나 존재할 때 그 원소는 W의 B 하한근사 $\underline{B}_\gamma(W)$에 속한다. X의 B 하한근사 $\underline{B}_\gamma(W)$에 속한 원소는 구간 오분류를 γ까지만 허용할 때 B의 관점에서 X의 멤버로 분류될 수 있다.

W의 B 상한근사 $\overline{B}_\gamma(W)$는 다음과 같이 정의한다.

$$\overline{B}_\gamma(W) = \{x \mid \exists y \in [x]_B \;(\exists w \in W (\forall b \in B \;(b(y) =_\gamma b(w))))\}$$

어떤 원소의 동치류에 있는 원소들 중 적어도 하나에 대하여 B의 관점에서 구간 오분류 허용 동치관계에 있는 원소가 W에 적어도 하나 존재할 때, 그 원소는 W의 B 상한근사 $\overline{B}_\gamma(W)$에 속한다. 상한근사 $\overline{B}_\gamma(W)$에 속한 원소는 구간 오분류를 γ까지만 허용할 때 W의 가능한 멤버(possible member)로 분류될 수 있다.

$BRB_\gamma(W) = \overline{B}_\gamma(W) - \underline{B}_\gamma(W)$는 구간 오분류 허용 동치관계에 대하여 W의 B-경계영역(B-boundary region)이라 부르며 B의 관점에서 구간 오분류를 γ까지만 허용할 때 X의 멤버로 분류될 수는 없는, 그러나 X의 가능한 멤버로 분류될 수 있는 원소들로 이루어져 있다.

$U - \overline{B}_\gamma(W)$는 W의 B-부정영역(B-negative region)이라 부르며, B에 의거하여 구간 오분류를 γ까지만 허용하는 확신을 가지고 X의 멤버가 아니라고 분류된다.

3. 구간 값을 갖는 가변정밀도 러프집합모형의 특성

오분류를 허용하지 않는 Pawlak의 러프집합 모형의 하한근사 및 상한근사에 관한 11가지 특성은 Ziarko의 가변정밀도 러프집합 모형에서는 그대로 유지된다[11]. 이제, 이러한 특성이 본 논문에서 제시하는 구간 오분류를 허용하는 가변정밀도 러프집합 모형에서도 유지되는지 살펴보자.

$\gamma \ne 1$인 구간 오분류를 허용하는 가변정밀도 러프집합 모형에서 Pawlak이 제시한 상한근사 및 하한근사에 관한 특성 식 (1)~식 (11)에 대응하는 특성의 성립여부는 다음과 같다.

$$\underline{B}_\gamma(W) \subseteq W \subseteq \overline{B}_\gamma(W) \tag{1}$$

$x \in \underline{B}_\gamma(W)$이면 하한근사의 정의에 의해 $[x]_B \subseteq W$가 된다. 그런데, $x \in [x]_B$이므로, $\underline{B}_\gamma(W) \subseteq W$는 항상 성립한다. $x \in W$이면 $\exists w \in W\,(\forall b \in B\;(b(x) =_\gamma b(w)))$가 항상 성립하므로 $W \subseteq \overline{B}_\gamma(W)$는 항상 성립한다.

$$\underline{B}_\gamma(\varnothing) = \overline{B}_\gamma(\varnothing) = \varnothing,\;\; \underline{B}_\gamma(U) = \overline{B}_\gamma(U) = U \tag{2}$$

$\gamma \ne 1$일 때 $\varnothing$에 대한 동치류는 $\varnothing$뿐이므로 $\underline{B}_\gamma(\varnothing) = \overline{B}_\gamma(\varnothing) = \varnothing$이 성립한다. $x \in U$이므로 $\exists u \in U\,(\forall b \in B\;(b(x) =_\gamma b(u)))$가 항상 성립하고, 이에 따라 $\underline{B}_\gamma(U) = \overline{B}_\gamma(U) = U$는 항상 성립한다.

$$\overline{B}_\gamma(V \cup W) = \overline{B}_\gamma(V) \cup \overline{B}_\gamma(W) \tag{3}$$

$[x]_B$의 어느 한 원소가 $V \cup W$의 한 원소와 구간 오분류 허용 동치관계가 있으면, 이 원소는 V의 한 원소와 구간 오분류 허용 동치관계가 있거나, 또는 W의 한 원소와 구간 오분류 허용 동치관계가 있으므로, 상한근사에 관한 위 식이 성립한다.

$$\underline{B}_\gamma(V \cap W) = \underline{B}_\gamma(V) \cap \underline{B}_\gamma(W) \tag{4}$$

$[x]_B$의 각 원소가 $V \cap W$의 한 원소와 구간 오분류 허용 동치관계가 있으면, 각 원소는 V의 한 원소와 구간 오분류 허용 동치관계가 있으면서, 동시에 W의 한 원소와 구간 오분류 허용 동치관계가 있게 된다. 만일 $[x]_B$의 한 원소가 V의 어느 원소와도 구간 오분류 허용 동치관계가 없거나 W의 어느 원소와도 구간 오분류 허용 동치관계가 없으면, 이 원소는 $V \cap W$의 어느 원소와도 구간 오분류 허용 동치관계가 없다. 따라서, 위 식이 성립한다.

$$V \subseteq W \text{이면 } \underline{B}_\gamma(V) \subseteq \underline{B}_\gamma(W) \text{이다.} \tag{5-1}$$

$[x]_B$의 각 원소가 V의 한 원소와 구간 오분류 허용 동치관계가 있으면, $V \subseteq W$인 W의 한 원소와도 구간 오분류 허용 동치관계가 존재한다. 따라서 위 식이 성립한다.

$$V \subseteq W \text{이면 } \overline{B}_\gamma(V) \subseteq \overline{B}_\gamma(W) \text{이다.} \tag{5-2}$$

$[x]_B$의 한 원소가 V의 한 원소와 구간 오분류 허용 동치관계가 있으면, $V \subseteq W$인 W의 한 원소와도 구간 오분류 허용 동치관계가 존재한다. 따라서 위 식이 성립한다.

$$\underline{B}_\gamma(V \cup W) \supseteq \underline{B}_\gamma(V) \cup \underline{B}_\gamma(W) \tag{6}$$

식 (5-1)로부터 $\underline{B}_\gamma(V \cup W) \supseteq \underline{B}_\gamma(V)$이고 $\underline{B}_\gamma(V \cup W) \supseteq \underline{B}_\gamma(W)$이므로 위 식이 성립한다.

$$\overline{B}_\gamma(V \cap W) \subseteq \overline{B}_\gamma(V) \cap \overline{B}_\gamma(W) \tag{7}$$

식 (5-2)로부터 $\overline{B}_\gamma(V \cap W) \subseteq \overline{B}_\gamma(V)$이고 $\overline{B}_\gamma(V \cap W) \subseteq \overline{B}_\gamma(W)$이므로 위 식이 성립한다.

$$\underline{B}_\gamma(-W) = -\overline{B}_\gamma(W) \tag{8}$$

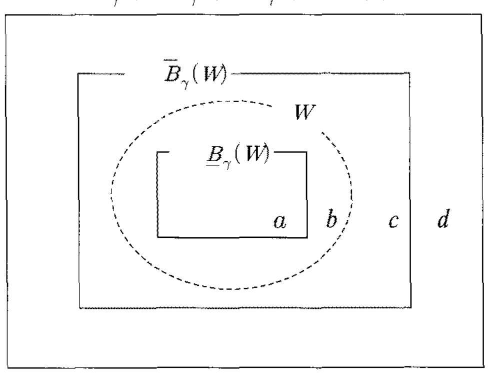
$[x]_B \subseteq (-W)$일 때에만 x가 $\underline{B}_\gamma(-W)$의 원소가 된다. $x \in \overline{B}_\gamma(W)$이면 $[x]_B$의 적어도 한 원소는$(-W)$의 어떤 한 원소와도 구간 오분류 허용 동치관계가 존재하지 않으므로, 그러한 x는 $\underline{B}_\gamma(-W)$의 원소가 될 수 없다. $x \notin \overline{B}_\gamma(W)$이면, $[x]_B$의 각 원소는 $(-W)$의 한 원소와 구간 오분류 허용 동치관계가 존재한다. 따라서, 위 식이 성립한다.

$$\overline{B}_\gamma(-W) = -\underline{B}_\gamma(W) \tag{9}$$

$[x]_B$의 적어도 한 원소가 $(-W)$의 어떤 한 원소와도 구간 오분류 허용 동치관계가 존재할 때, $x \in \overline{B}_\gamma(-W)$가 된다. $[x]_B \subseteq W$이면 당연히 $x \notin \overline{B}_\gamma(W)$이다. $[x]_B \subseteq W$가 성립하지 않으면, $[x]_B$의 적어도 한 원소가 $(-W)$의 어떤 한 원소와도 구간 오분류 허용 동치관계가 존재한다. 따라서, 위 식이 성립한다.

아래 특성들의 성립여부를 설명하기 위하여 다음과 같은 경우를 가정하자(<그림 1> 참조).

$$a =_\gamma b,\ b =_\gamma c,\ c =_\gamma d,\ W = \{a, b\}$$



<그림 1> 러프집합

그러면,
$[a]_B = \{a, b\}$, $[b]_B = \{a, b, c\}$, $[c]_B = \{b, c, d\}$, $[d]_B = \{c, d\}$가 되고, 아울러 $\underline{B}_\gamma(W) = \{a\}$, $\overline{B}_\gamma(W) = \{a, b, c\}$가 된다.

$$\underline{B}_\gamma(\underline{B}_\gamma(W)) = \overline{B}_\gamma(\underline{B}_\gamma(W)) = \underline{B}_\gamma(W) \tag{10}$$

$$\begin{aligned} \overline{B}_\gamma(\underline{B}_\gamma(W)) &= \overline{B}_\gamma(\{a\}) \\ &= \{a, b\} \\ &\neq \underline{B}_\gamma(W) \end{aligned}$$

따라서, 위 식은 성립하지 않는다.

$$\overline{B}_\gamma(\overline{B}_\gamma(W)) = \underline{B}_\gamma(\overline{B}_\gamma(W)) = \overline{B}_\gamma(W) \quad (11)$$

$$\begin{aligned}\underline{B}_\gamma(\overline{B}_\gamma(W)) &= \underline{B}_\gamma(\{a, b, c\}) \\ &= \{a, b\} \\ &\neq \underline{B}_\gamma(W)\end{aligned}$$

따라서, 위 식은 성립하지 않는다.

4. 결론 및 향후 연구과제

본 논문에서는, 객체의 속성 값이 구간을 가지는 경우, 두 객체의 속성 값이 같은지를 판단할 때 구간 오분류를 허용하는 가변정밀도 러프집합모형을 제시하였다. 이를 위하여, 동치류를 정의하였고, 하한근사 및 상한근사를 정의하였으며, Pawlak[6]이 제시한 근사의 11가지 특성이 본 모형에서도 성립하는지를 검증하였다. 근사 특성 식 (1)~식 (9)는 성립하였으나, 식 (10)과 식 (11)은 성립되지 않음을 예를 통하여 보였다.

향후 연구과제로는, 구간 값을 갖는 데이터에 대한 클러스터링 알고리듬과 분류 알고리듬에서 본 논문에서 제시한 모형이 어떻게 응용될 수 있는지와, 구간 오분류 값과 결과와의 상관관계에 관한 연구 등이 있다.

참고문헌

[1] Grzymala-Busse, J. W. and Ziarko, W.; "Data mining based on rough sets," (ed.) J. Wang, Data Mining : Opportunities and Challenges, *Idea Group Publishing*, 142-173, 2003.

[2] Huang, K. Y.; "Application of VPRS Model with Enhanced Threshold Parameter Selection Mechanism to Automatic Stock Market Forecasting and Portfolio Selection," *Expert Systems with Applications*, 36(9) : 11652-11661, 2009.

[3] Li, W. H., Chen, S. B., and Wang, B.; "A Variable Precision Rough Set Based Modeling Method for Pulsed GTAW," *International Journal of Advanced Manufacturing Technology*, 36(11-12) : 1072-1079, 2008.

[4] Ningler, M., Stockmanns, G., Schneider, G., Kochs, H. D., and Kochs, B.; "Adapted Variable Precision Rough Set Approach for EEG Analysis," *Artificial Intelligence in Medicine*, 47(3) : 239-261, 2009.

[5] Pan, X., Zhang, S. Q., Zhang, H. Q., Na, X. D., and Li, X. F.; "A Variable Precision Rough Set Approach to the Remote Sensing and Use/Cover Classification," *Computers and Geoscience*, 36(12) : 1466-1473, 2010.

[6] Pawlak, Z.; "Rough sets," *International Journal of Information Science*, 11(5) : 341-356, 1982.

[7] Xie, G., Zhang, J. L., and Lai, K. K.; "Using VPRS to Mine the Significance of Risk Factors in IT Project Management," Proceedings of Rough Sets and Knowledge technology, *Lecture Notes in Artificial Intelligence 4062*, Springer Verlag : 750-757, 2006.

[8] Xie, G., Zhang, I. L., Lai, K. K., and Yu, L.; "Variable Precision Rough Set for Group Decision-Making : An Application," *International Journal of Approximate Reasoning*, 49(2) : 331-343, 2008.

[9] Xie, G., Yue, W. Y., Wang, S. Y., and Lai, K. K.; "Dynamic Risk Management in Petroleum Project Investment Based on a Variable Precision Rough Set Model," *Technological Forecasting and Social Change*, 77(6) : 891-901, 2010.

[10] Zhao, W. Q. and Zhu, Y. L.; "Classifying email Using Variable Precision Rough Set Approach," Proceedings of Rough Sets and Knowledge technology, *Lecture Notes in Artificial Intelligence 4062*, Springer Verlag : 766-771, 2006.

[11] Ziarko, W.; "Variable precision rough set model," *Journal of Computer and System Science*, 46 : 39-56, 1993.