http://dx.doi.org/10.14400/JDC.2014.12.6.223

러프집합 이론을 이용한 러프 엔트로피 기반 지식감축

박인규
중부대학교 컴퓨터학과

Rough Entropy-based Knowledge Reduction using Rough Set Theory

In-Kyoo Park
Dept. of Computer Science Joongbu University

요 약 대용량의 지식베이스 시스템에서 유용한 정보를 추출하여 효율적인 의사결정을 수행하기 위해서는 정제된 특징추출이 필수적이고 중요한 부분이다. 러프집합이론에 있어서 최적의 리덕트의 추출과 효율적인 객체의 분류에 대한 문제점을 극복하고 자, 본 연구에서는 조건 및 결정속성의 효율적인 특징추출을 위한 러프엔트로피 기반 퀵리덕트 알고리듬을 제안한다. 제안된 알고리듬에 의해 유용한 특징을 추출하기 위한 조건부 정보엔트로피를 정의하여 중요한 특징들을 분류하는 과정을 기술한다. 또한 본 연구의 적용사례로써 실제로 UCI의 5개의 데이터에 적용하여 특징을 추출하는 시뮬레이션을 통하여 본 연구의 모델링이 기존의 방법과 비교결과, 제안된 방법이 효율성이 있음을 보인다.

주제어 : 데이터 마이닝, 러프집합, 특징추출, 퀵리덕트, 러프 엔트로피

Abstract In an attempt to retrieve useful information for an efficient decision in the large knowledge system, it is generally necessary and important for a refined feature selection. Rough set has difficulty in generating optimal reducts and classifying boundary objects. In this paper, we propose quick reduction algorithm generating optimal features by rough entropy analysis for condition and decision attributes to improve these restrictions. We define a new conditional information entropy for efficient feature extraction and describe procedure of feature selection to classify the significance of features. Through the simulation of 5 datasets from UCI storage, we compare our feature selection approach based on rough set theory with the other selection theories. As the result, our modeling method is more efficient than the previous theories in classification accuracy for feature selection.

Key Words : Data Mining, Rough Set, Feature Selection, Quick-Reduct, Rough Entropy

1. 서론

지식 베이스를 복구하는 경우에 특징추출(feature selection)은 중요한 문제이다[1]. 일반적으로 중복되는 특징이 너무 많고 부적절한 특징들로 인하여 데이터를 판별하는데 중요한 특징이 작용하지 않을 수 있다. 따라

Received 3 March 2014, Revised 12 April 2014
Accepted 20 June 2014
Corresponding Author: In-Kyoo Park(Joongbu Univ.)
Email: fip2441g@gmail.com

ISSN: 1738-1916

서 이러한 문제에 대안으로 특징추출은 판별의 정확성을 높일 수도 있고 계산의 오버헤드를 줄일 수 있다. 적절한 방법을 이용하여 데이터베이스의 많은 차원을 줄일 수 있다. 특징추출의 주요한 목적은 원래의 특징을 나타내는데 어느 정도의 정확성을 유지하면서 해당 문제에 대하여 최소한의 특징을 결정하는 것이다. 일반적으로 특징추출을 위하여 러프집합의 활용방법은 hill-climbing와 확률적(stochastic)방법이다[2]. 전자는 속성의 중요도를 척도로 하여 속성을 평가하여 시작상태가 공집합이거나 코어에서 출발하여 속성을 추가하여 리덕트를 구성하는 전방선택(forward selection)과 달리 후방제거(backward elimination)는 중요하지 않은 속성을 제거해 나가는 방식이다. 속성을 평가하는 방법에는 크게 긍정영역을 이용하는 방법과 조건부 엔트로피를 이용하는 방법이 있다. 결론적으로 이러한 방법들은 가장 중요한 속성을 기반으로 데이터를 감축한다[3,4]. 한편으로 확률적인 방법은 유전자 알고리즘과 같은 방법을 러프집합 이론과 결합하여 수행하기 때문에 오랜 시간이 걸리고 찾아진 해의 보장이 어렵다는 단점이 있다[5]. 본 논문에서는 최적의 특징과 관련 지식 내에서 존재하는 가장 기본적인 개념을 정의하기에 충분한 지식의 필수적인 부분을 탐색하기 위하여 특징들의 상호확률을 고려하여 속성의 중요도에 대한 변별력을 향상시키는 척도를 정의하여 새로운 조건부 정보엔트로피를 제시하고자 한다.

2. 러프집합

지식은 임의의 영역내의 다양한 분류패턴들의 집합으로 구성되어 진다. U≠∅이 전체집합인 유한집합이고 R이 U의 동치관계들의 집합일 때 지식기반은 관계 시스템 K=(U, R)에 해당한다. P⊆R이고 P≠∅이라면 P에 해당하는 모든 동치관계들의 교집합도 역시 동치관계가 되고 이를 IND(P)라고 표기하고 다음과 같이 P의 식별불가능 관계(indiscernibility relation over P)라고 한다[6].

$$[x]_{IND(P)} = \cap_{R \in P}[x]_R \tag{1}$$

따라서 U/IND(P)는 K내의 U에 관한 P의 집합과 관련된 지식을 나타낸다. 러프집합에서는 임의의 지식을 구성하는 개념들간의 의존성을 기반으로 데이터에 포함되어 있는 불필요한 동치관계를 제거할 수 있다.

Xi⊆U인 집합들에 대하여 F={X1,...,Xn}을 고려해 보자. ∩(F-{Xi})≠∩F이면 집합 Xi는 F내에서 필요 불가결하다고 하고 그렇지 않으면 Xi는 F내에서 불필요하다고 한다. 러프집합은 동치관계에 의한 granularity의 기반에서 하한근사와 상한근사라고 하는 개념을 이용하여 효율적인 분류(classification)를 수행한다. X⊆U와 동치관계 R∈IND(K)를 써서 두 개의 집합 R의 하한근사와 R의 상한근사를 다음과 같이 정의 할 수 있다.

$$\begin{aligned} \underline{R}X &= \{x \in U | [x]_R \subseteq X\} \\ \overline{R}X &= \{x \in U | [x]_R \cap X \neq \varnothing\} \end{aligned} \tag{2}$$

집합 BNDR(X)=$\overline{R}$X-$\underline{R}$X는 X의 R경계라고 하고 집합 $\underline{R}$X는 지식 R내에서 X의 원소로 확실하게 분류되는 U의 모든 원소들의 집합이고, $\overline{R}$X는 분류될 가능성이 있는 U의 원소들의 집합이며 BNDR(X)는 지식 R로써 X또는 -X의 어느 집합에도 분류될 수 없는 원소들의 집합으로서 다음과 같다.

$$\begin{aligned} POS_R(X) &= \underline{R}X \\ NEG_R(X) &= U - \overline{R}X \\ BND_R(X) &= \overline{R}X - \underline{R}X \end{aligned} \tag{3}$$

P와 Q를 U의 동치관계라고 할 경우에 Q가 P에 k(0≤k≤1)정도로 의존하고 있다는 것을 P⇒kQ라고 나타내고 다음과 같이 정의된다.

$$k = \gamma_P(Q) = \frac{|POS_P(Q)|}{|U|} \tag{4}$$

$\gamma_P(Q)$는 분류의 질(quality)을 나타낸다. k가 1이면 Q는 P에 완전히 의존하고 k가 0이면 Q는 P에 의존하지 않으며 나머지의 경우는 부분적으로 의존한다. Q의 P-긍정영역은 다음과 같이 정의 할 수 있고 U/P에 의해 U/Q의 동치류에 분류될 수 있는 U의 모든 객체들의 집합이다.

$$POS_P(Q) = \bigcup_{X \in U/Q} \underline{P}X \quad (5)$$

S⊆P가 P의 독립(independent)이고 POSS(Q) = POSP(Q)이면 S는 P의 Q-리덕트(Q-reduct)라고 한다. 이와 같이 리덕트를 이용하여 중복되는 속성을 제거하여 감축된 속성만으로 원래의 속성에 의한 분류와 동일한 결과를 도출할 수 있다.

3. 지식의 감축

일반적으로 특징추출 방법은 평가함수나 임의의 척도를 이용하여 여러 가지의 특징을 평가하여 데이터가 분류되는 부류와 관계가 있는 특징들을 추출하는 방법이다. 실제적으로 특징추출을 통하여 데이터에 포함되어 있는 잡음과 관련성이 없는 특징과 유해한 특징을 제거할 수 있기 때문에 중요하다. 이와 같이 추출된 특징의 유용성은 결정특징(decision feature)을 예견할 수 있는 관련성이나 다른 특징과 밀접하게 관련되어 있는 중복성에 의하여 평가된다.

3.1 퀵리덕트

탐색은 모든 경우를 고려하지 않고 처음에는 공집합에서 출발하여 원래의 데이터에 대한 의존성의 값과 동일해 질 때 까지 순차적으로 하나씩 속성을 추가해 나간다[7]. 이러한 속성들은 러프집합에 의한 임의의 의존성의 척도에 의해 가장 확실한 값을 가지고 각각의 속성의 의존도에 의해 최선의 후보속성을 결정한다.

```
QR(C, D)
C: the set of all conditional features
D: the set of all decision features
    (1) R ← { }
    (2) do
    (3)    T←R
    (4)    ∀x∈(C-R)
    (5)    if K_{R∪{x}}(D) > K_T(D)
    (6)       T←R∪{x}
    (7)    R←T
    (8) until K_R(D) = K_C(D)
    (9) return R
```

[Fig. 1] QuickReduct Algorithm

3.2 엔트로피 기반 리덕트

엔트로피의 척도를 이용하여 속성들의 등급을 결정하며 임의의 종료조건의 임계값이 필요하지 않다.

```
EBR(C,D)
C: the set of all consitional features
D: the set of all decision features
    (1) R ← { }
    (2) do
    (3)    T←R
    (4)    ∀x∈(C-R)
    (5)    if H(R∪{x}) > H(T)
    (6)       T←R∪{x}
    (7)    R←T
    (8) until H(D|R) = H(D|C)
    (9) return R
```

[Fig. 2] Entropy QuickReduct Algorithm

원래의 속성에 의한 엔트로피 값과 동일한 결과를 도출하는 속성의 엔트로피 값이 동일하면 알고리즘이 종료하게 된다[8].

3.3 관계 리덕트

관계 의존성(relative dependency)의 척도를 기반으로 식별함수나 긍정영역의 계산을 피하기 위한 방법이지만 최적화를 보장하지는 않는다. 따라서 기존의 러프집합에 의한 의존성의 척도를 관계 의존성이라는 다른 방법으로 대체하였다[9,10].

```
RR(C,D)
C: the set of all consitional features
D: the set of all decision features
    (1) R ← C
    (2) ∀a∈C
    (3)    if K_{R∪{a}}(D) = 1)
    (4)       R←R-{a}
    (5) return R
```

[Fig. 3] Relative QuickReduct Algorithm

일반적으로 데이터 마이닝에서 결정부류가 알려지지 않을 수 있고 정보가 부족할 수 있는 경우에 자율적인 특징추출방법이 필요하다. 기존의 알고리즘은 조건부와 결정부의 입력을 가지는 기존의 방법과 달리 조건부 속성만으로 구성되어 속성간의 의존성의 정도에 의해 중요속

성을 도출한다.

3.4 자율학습 리덕트

속성이 가지는 모든 경우의 수를 고려하지 않고 집합의 긍정영역을 기반으로 결정부의 기대치를 가지지 않고 평가척도를 수행한다.

$$\gamma_P(a) = |POS_P(a)|/|U|, \ \forall a \in A \quad (6)$$

```
USQR(C)
C: the set of all conditional features
    (1) R ← { }
    (2) do
    (3) T←R
    (4)    ∀x∈(C-R)
    (5)    ∀y∈C
    (6)    r_{R∪{x}}(y) = |POS|_{R∪{x}}(y) /|U|
    (5)    if r_{R∪{x}}(D), ∀y∈C > r_T(D), ∀y∈C
    (6)    T←R∪{x}
    (7) R←T
    (8) until r_R(y) = r_C(y)
    (9) return R
```

[Fig. 4] Relative QuickReduct Algorithm

처음에는 공집합에서 출발하여 원래의 데이터와 동일한 의존성을 가질 경우에 순차적으로 속성을 추가한다. 따라서 하나의 속성에 대한 의존도를 계산하여 각각의 조건부 속성을 구성하고 모든 조건부 속성의 평균 의존도의 산출을 통하여 최적의 후보속성이 선택되어 진다.

4. 러프 엔트로피기반 감축

S=(U,C,D)에서 U/C={C1,C2,.,Ck}, U/D={D1,D2,., Dn}이라고 하자. 제안된 방법에서는 조건부 속성만을 이용하여 특징을 추출하기 때문에 조건속성 Ci에 대하여 Cj의 종속성은 조건부 정보엔트로피를 이용하여 식(6)에서와 같이 두 속성의 확률을 동시에 고려하여 속성의 변별력을 향상시켰다.

$$H(C_j|C_i) = -\sum_{i=1}^{m}\frac{|C_i|}{|U|}\sum_{j=1}^{n}\frac{|C_j|}{|U|} ln\frac{|C_j \cap C_i|}{|C_i|} \quad (7)$$

종속성은 부분적으로 지식 C_j의 일부분이 지식 C_i로부터 추출이 가능함을 의미한다. 따라서 이를 정보엔트로피 개념에 의해 식(7)과 같이 정의할 수 있다. 알고리즘에서 처음에는 공집합에서 출발하여 원래의 데이터에 대한 의존성이 가장 큰 속성을 추출한다. 각 속성집합의 종속성의 평균을 산출하여 최적의 속성을 선택하게 된다.

$$K_{C_i}(C_j) = |1 - H(C_j|C_i)|/|U| \quad (8)$$

〈Table 1〉 a knowledge system

$x \in U$	a	b	c	d
1	1	0	2	1
2	1	0	2	0
3	1	2	0	0
4	1	2	2	1
5	2	1	0	0
6	2	1	1	0
7	2	1	2	1

```
EBR(C,D)
C: the set of all consitional features
D: the set of all decision features
    (1) R ← { }
    (2) do
    (3) T←R
    (4)    ∀x∈(C-R)
    (5)    ∀y∈C
    (6)    K_{R∪{x}}(y)=|1-H(R∪{x}|y)|/|U|
    (5)    if K_{R∪x}(D),∀y∈C > K_T(D), ∀y∈C
    (6)        T←R∪{x}
    (7) R←T
    (8) until K_R(y) = K_C(y)
    (9) return R
```

[Fig. 5] Rough Entropy QuickReduct Algorithm

제안된 알고리즘을 검증하기 위하여 <Table 1>에서 주어진 정보시스템을 고려해 보자. 조건부 속성은 {a,b,c,d}이고 각각의 속성의 의존도는 다음과 같다.

H({1,2,3,4}|{1,2,3,4})=-4/7*4/7ln(4/4)=0
H({1,2,3,4}|{5,6,7})=-3/7*4/7ln(0/3)=0.245
H({5,6,7}|{1,2,3,4})=-4/7*3/7ln(0/4)=0.245
H({5,6,7}|{5,6,7})=-3/7*3/7ln(3/3)=0
H({b}|{1,2,3,4})=min(0,0.245)=0

H({b}|{5,6,7})=min(0,0.245)=0
H({b}|{a})=mean(0,0)=0

H({1,2}|{1,2,3,4})=-4/7*2/7ln(2/4)=0.113
H({1,2}|{5,6,7})=-3/7*2/7ln(0/3)=0.122
H({3,4}|{1,2,3,4})=-4/7*2/7ln(2/4)=0.113
H({3,4}|{5,6,7})=-3/7*2/7ln(0/3)=0.122
H({5,6,7}|{1,2,3,4})=-4/7*3/7ln(0/4)=0.184
H({5,6,7}|{5,6,7})=-3/7*3/7ln(3/3)=0
H({b}|{1,2,3,4})=min(0.113,0.113,0.184)=0.113
H({b}|{5,6,7})=min(0.122,0.122,0)=0
H({b}|{a})=mean(0.113,0)=0.0565

H({1,2,4,7}|{1,2,3,4})=-4/7*4/7ln(3/4)=0.094
H({1,2,4,7}|{5,6,7})=-3/7*4/7ln(1/3)=0.269
H({3,5}|{1,2,3,4})=-4/7*2/7ln(1/4)=0.226
H({3,5}|{5,6,7})=-3/7*2/7ln(1/3)=0.135
H({6}|{1,2,3,4})=-4/7*1/7ln(0/4)=0.082
H({6}|{5,6,7})=-3/7*1/7ln(1/3)=0.067
H({b}|{1,2,3,4})=min(0.094,0.226,0.082)=0.082
H({b}|{5,6,7})=min(0.269,0.135,0.067)=0.067
H({b}|{a})=mean(0.082,0.067)=0.0745

H({1,4,7}|{1,2,3,4})=-4/7*3/7ln(2/4)=0.170
H({1,4,7}|{5,6,7})=-3/7*3/7ln(1/3)=0.202
H({2,3,5,6}|{1,2,3,4})=-4/7*4/7ln(2/4)=0.226
H({2,3,5,6}|{5,6,7})=-3/7*4/7ln(2/3)=0.099
H({b}|{1,2,3,4})=min(0.170,0.226)=0.170
H({b}|{5,6,7})=min(0.202,0.099)=0.099
H({b}|{a})=mean(0.170,0.099)=0.1345

H({a,b,c,d}|{a})=mean(0+0.0565+0.0745+0.1345)/4
=0.066375

〈Table 2〉 Degree of Attribute by Positive Region

y|x	{a}	{b}	{c}	{d}
a	1.0	1.0	0.1429	0.0
b	0.4286	1.0	0.1429	0.0
c	0.0	0.2857	1.0	0.4286
d	0.0	0.0	0.4286	1.0
K{P}(y), ∀y∈C	0.3571	0.5714	0.4285	0.3571

〈Table 3〉 Degree of Attribute by Rough Entropy

y|x	{a}	{b}	{c}	{d}
a	0	0	0.0199	0.0485
b	0.0566	0.0283	0.0377	0.0663
c	0.0806	0.0609	0.0406	0.0446
d	0.1345	0.1121	0.0826	0.0619
K{P}(y), ∀y∈C	0.0679	0.0503	0.0452	0.0553

〈Table 4〉 Degree of Attribute by Positive Region

y|x	{a,b}	{b,c}	{b,d}
a	1.0	1.0	1.0
b	1.0	1.0	1.0
c	0.2857	1.0	0.7143
d	0.0	0.7143	1.0
K{P}(y), ∀y∈C	0.5714	0.9286	0.9286

비슷한 방법으로 다른 속성들간의 의존도가 <Table 2>에 나타나 있다. 속성 {c}는 가장 높은 의존도를 나타내기 때문에 분할리덕트 속성으로 설정되어 최적의 리덕트의 후보집합은 {a,c}, {b,c}, {c,d}로 구성되어고 각각의 리덕트에 대한 의존도를 계산하면 <Table 3>과 같다.

〈Table 5〉 Degree of Attribute by Rough Entropy

y|x	{a,c}	{b,c}	{c,d}
a	0	0	0.0153
b	0.0099	0.0050	0.0192
c	0	0	0
d	0.0149	0.0145	0.0097
K{P}(y), ∀y∈C	0.0062	0.0049	0.0111

<Table 3>의 각각의 리덕트에 대하여 의존도를 계산하면 <Table 4>와 같다. 역시 가장 높은 의존도를 나타내는 리덕트는 {b,c}이므로 분할 리덕트로 설정되어 진다. 두 개의 리덕트 후보중에서 {b,c,d} 리덕트가 더 높은 의존도를 가지기 때문에 최종적인 리덕트로 설정되어 원래의 속성에 의한 의존도와 가장 근접한 의존도를 나타낸다.

〈Table 6〉 Degree of Attribute by Positive Region

y|x	{a,b,c}	{b,c,d}
a	1.0	1.0
b	1.0	1.0
c	1.0	1.0
d	0.7143	1.0
K{P}(y), ∀y∈C	0.9285	1.0

〈Table 7〉 Degree of Attribute by Rough Entropy

y|x	{a,b,c}	{b,c,d}
a	0	0
b	0	0
c	0	0
d	0.0141	0.0071
K{P}(y), ∀y∈C	0.0035	0.0018

제안한 방법과 기존의 USQR에 의한 방법은 동일한 결과를 가진다. 차이점은 기존의 방법은 리덕트에 의한 의존도가 원래의 의존도와 동일하지만, 제안한 방법에서는 원래의 의존도에 가장 근접한 값을 가지는 리덕트를 설정하게 된다. 기존방법에 의한 {b,c}, {b,d}의 리덕트의 의존도는 0.92857로 동일한 의존도를 나타내었다. 그러나 제안된 방법에서는 동일한 의존도를 가지는 리덕트가 발생하지 않기 때문에 속성간의 특징을 추출하는데 따른 변별력을 높여준다고 할 수 있다.

5. 실험 및 결과고찰

WEKA는 자바기반의 유용한 데이터 마이닝 도구이다. 본 논문에서는 WEKA에서 구현되어 있는 네 가지의 벤치마크 분류기를 이용하여 QR, EBR, RR의 기존 알고리즘과 제안한 알고리즘의 효용성을 실험하였다. 논문에 사용된 분류기는 DTNB, JRip, J48과 LMT이다. 실험에 사용된 모든 데이터는 UCI 데이터베이스의 자료를 이용하였다. 동일한 자료에 대하여 여러 가지의 방법들의 특징추출, 특징추출 시간과 분류의 정확도에 대하여 비교하였다. 벤치마크용 데이터는 <Table 5>와 같다.

〈Table 8〉 Datasets

no.	Datasets	objects	classes	attributes
1	Iris	150	3	4
2	WBCD	699	2	9
3	BUPAiver	345	2	6
4	Pimaln Diabetes	768	2	8
5	Wine	178	3	13

〈Table 9〉 Selected features

no.	QR	EBR	RR	USQR	RER
1	1,2,3	1,2,3	2,3,4	1,2,3,4	1,2,3
2	1,2,6,7	1,2,6,7	5,6,7,8,9	1,...,9	1,2,3,4,5,6,7,8
3	1,2,5	1,2,5	3,4,5	2,3,5	2,3,4
4	1,2,7	1,2,7	6,7,8	1,2,7	1,2,3,4,6
5	1,7	1,7	12,13	1,2	1,7,10,12

다섯 가지의 데이터에 대하여 각각의 방법에 의한 감축된 특징이 <Table 9>에 나타나 있고 이러한 특징에 의한 분류의 정확도를 <Table 10>에 나타내었다. 감축되지 않은 특징에 대하여 감축된 특징에 의한 분류의 정확도가 일부의 분류기에서 상승한 것을 알 수 있었다. 이러한 상승의 내용을 데이터 별로 보면 iris의 경우 RR, 제안된 방법이 각각 3개와 2개의 분류기에서 상승하였고, WBCD의 경우 제안된 방법, QR과 EBR이 3개의 분류기에서 동일하게 상승하였고, BUPALiver의 경우 RR이 2개의 분류기에서 그리고 다른 방법들은 1개의 분류기에서 상승하였다.

〈Table 10〉 Classification accuracy values

Algorithm	Classiifers	Data sets				
		1	2	3	4	5
Unreduced data	DTNB	92.00	96.85	57.68	73.82	98.88
	JRip	94.00	95.42	64.63	76.04	92.13
	J48	96.00	94.56	68.69	73.82	93.82
	LMT	94.00	95.99	66.37	77.47	97.19
QR	DTNB	94.00	97.28	57.68	73.31	92.13
	JRip	92.67	95.70	62.32	71.88	90.44
	J48	93.33	95.56	92.90	72.14	94.94
	LMT	93.33	95.28	60.87	75.00	90.45
EBR	DTNB	94.00	97.28	57.68	73.31	92.13
	JRip	92.67	95.70	62.32	71.88	90.44
	J48	93.33	95.56	92.90	72.14	94.94
	LMT	93.33	95.28	60.87	75.00	90.45
RR	DTNB	62.67	96.13	57.68	67.18	86.52
	JRip	93.33	94.13	64.92	68.61	87.64
	J48	96.00	94.27	63.76	67.70	89.89
	LMT	95.33	95.42	64.63	69.92	89.89
USQR	DTNB	92.00	96.85	57.68	73.31	78.65
	JRip	94.00	95.42	63.77	71.88	78.65
	J48	96.00	94.56	66.96	72.14	78.09
	LMT	94.00	95.99	65.50	75.00	80.34
RER	DTNB	94.00	97.14	57.97	74.08	93.26
	JRip	92.67	94.71	63.19	74.47	92.70
	J48	93.33	94.85	57.10	73.96	96.63
	LMT	93.33	96.14	64.35	76.56	91.57

또한 Pimaln Diabetes과 Wine의 경우 제안된 방법이 2개의 분류기에서 상승하였다. USQR의 경우 iris에서 원 데이터와 동일한 결과를 나타내었지만 감축된 특징이 다른 방법보다 많았고 WBCD에서는 감축이 이루어지지 않았다. 전반적으로 제안된 방법이 다섯 가지의 데이터에 대하여 세 가지 이상에서 우세함을 나타내었다.

6. 결론

본 논문에서는 러프집합과 러프 엔트로피를 이용하여 특징을 추출하는 퀵리덕트 알고리즘을 제안하였다. 제안된 방법에서는 리덕트의 모든 경우의 수를 발생하지 않고 리덕트를 탐색한다. 러프 엔트로피개념을 가변 러프 집합모형에 적용하여 속성들간의 의존도의 변별력을 고려하였다. 그 결과 기존의 방법보다 속성들간의 의존도의 중복성을 피할 수 있었다. WEKA에 의한 분류기의 실험을 통하여 분류능력과 분류오차에서 기존의 방법보다 중복속성을 효과적으로 제거할 수 있음을 알 수 있었다. 향후에 대용량 지식베이스에 적용하여 제안된 방법의 안정성을 확보해야 할 것으로 사료된다.

REFERENCES

[1] J. W. Grzymala-Busse, "LERS-a system for learning from examples based on rough sets", in Intellegent Decision Support, Kruwer Academic Publishers, pp. 3-18, 1992.

[2] M. Dashand and H Liu, "Feature selection for classification", Intelligent Data Analysis, Vol. 1, No. 3, pp. 131-156, 1997.

[3] M. Dash and H. Liu, "Unsupervised feature selction", in Proc. of the Pacific and Asia Conf. on Knowledge Discovery and Data Mining, Kyoto, pp. 110-121, 2000.

[4] C. Velayutham and K. Thangaval, "Unsupervised Quick Reduct Algorithm using Rough Set Theory", Jouranl of Electronic Seience and Technology Vol. 9, No. 3, pp. 193-201, 2011.

[5] S. K. Das, "Feature selection with a linear dependence measure", IEEE Trans. on Computers, Vol. 20, No. 9, pp. 1106-1109, 1971.

[6] Lin Sun, "Decision Table Reduction Method Based on New Conditional Entropy for Rough Set Theory", International Workshop on Intelligent Systems and Applications, pp. 23-24, May 2009

[7] Baoxiang Liu, Ying Li, Lihong Li, Yaping Yu, "An Approximate Reduction Algorithm Based on Conditional Entropy", Information Computing and Applications, Vol. 106, pp. 319-32, 2010

[8] Zhangyan Xu, Jianhua Zhou, Chenguang Zhang, "A Quick Attribute Reduction Algorithm Based on Incomplete Decision Table", Information Computing and Applications, Vol. 391, pp. 499-508, 2013

[9] K. Thankavel and A. Pethalakshmi, "Dimensionality reduction based on rough set theory: a review", Applied Soft Computing, Vol. 9, No. 1, pp.1-12, 2009.

[10] J. Han, X. Hu and T.-Y. Lin, "Feature sebset selection based on relative dependency between attributes", in Proc. of the 4th International Conf. on Rough Sets and Current Trends in Computing, Uppsala, pp. 176-185, 2004.

박 인 규(Park, In Kyoo)

- 1985년 2월 : 연세대학교 전지과(공학석사)
- 1997년 2월 : 원광대학교 전자과(공학박사)
- 1997년 3월 ~ 현재 : 중부대학교 컴퓨터학과 교수
- 관심분야 : 러프집합, 퍼지집합
- E-Mail : fip2441g@gmail.com