Regular paper http://dx.doi.org/10.13067/JKIECS.2015.10.6.683

# 속성 변동 최소화에 의한 러프집합 누락 패턴 부합

이영천*

# Missing Pattern Matching of Rough Set Based on Attribute Variations Minimization in Rough Set

Young-Cheon Lee*

요약

러프집합에서 누락된 속성 값들은 Reduct와 Core 계산, 더 나아가서 결정 트리 구축에 있어서 식별 불능의 패턴 부합 문제를 가진다. 현재 누락된 속성 값들의 추정과 관련하여 보편적인 속성 값으로의 대체, 속성들의 모든 가능한 값 할당, 이벤트 포장 방법, C4.5, 특수한 LEM2 알고리즘과 같은 접근방식들이 적용되고 있다. 그렇지만, 이들 접근방식은 결국 전형적으로 자주 등장하는 속성 값 혹은 가장 보편적인 속성 값으로의 단순 대체를 나타내기 때문에, 주요 속성 값들이 누락된 경우에 정보 손실이 큰 의사 결정 규칙들이 유도되기 때문에 의사결정 규칙들의 교차 검증에서 문제가 된다. 본 연구에서는 이러한 문제점을 개선시키기 위해 속성들 간에 엔트로피 변동을 활용하여 정보 이득이 높은 방향으로 누락된 속성 값들을 대체하는 방식을 제안한다. 제안된 접근방식에 관한 타당성 검토는 비교적 가까운 유사 관계에 의해 누락 값 대체 방식을 적용하는 ROSE 프로그램과의 비교를 나타낸다.

ABSTRACT

In Rough set, attribute missing values have several problems such as reduct and core estimation. Further, they do not give some discernable pattern for decision tree construction. Now, there are several methods such as substitutions of typical attribute values, assignment of every possible value, event covering, C4.5 and special LEMS algorithm. However, they are mainly substitutions into frequently appearing values or common attribute ones. Thus, decision rules with high information loss are derived in case that important attribute values are missing in pattern matching. In particular, there is difficult to implement cross validation of the decision rules. In this paper we suggest new method for substituting the missing attribute values into high information gain by using entropy variation among given attributes, and thereby completing the information table. The suggested method is validated by conducting the same rough set analysis on the incomplete information system using the software ROSE.

키워드

## Ⅰ. 서론

러프집합 이론은 불완전 정보 테이블을 바탕으로 하기 때문에 누락된 속성 값들의 처리는 추후 의사결정 규칙들의 구축과 관련하여 매우 민감한 문제가 된다. 표준 러프집합 이론의 측면에서 하한 및 상한 근사는

* 교신저자(corresponding author) : 호남대학교 컴퓨터공학과(yclee@honam.ac.kr)
접수일자 : 2015. 05. 12 심사(수정)일자 : 2015. 06. 13 게재확정일자 : 2015. 06. 23

보통집합들이라 할 수 있지만, 다른 관점에서 보면 퍼지집합일 수도 있다. 속성 값이 누락되는 이유는 기본적으로 두 가지가 있는데, 원래 속성 값이 존재하나 어떠한 이유로 인해 삭제된 경우에 Grzymala- Busse는 누락된 속성 값들을 제외한 상황에서 식별 불능 집합을 유지하면서 보편적인 속성 값으로 대체하는 방식을 소개하였다. 부적절한 이유로 인해 원래 값들이 기록되지 않은 경우 또한 Grzymala-Busse는 러프집합의 상한 근사에 속하는 빈도가 가장 높은 속성 값으로의 대체가 적절함을 보였다. 문제는 이러한 접근방식에 의한 누락 값 대체가 결정 트리 구축 알고리즘에 따라 상이한 결과들이 나타나고 있다는 점이다[4].

일반적으로 불완전한 의사결정표는 식별 불능 관계로 기술되는 완전 의사결정표에서와 유사하게 특징적인 관계들로 기술된다. 러프집합 이론에서 완전 의사결정표에 관해 일단 식별 불능 관계가 고정되고 사례집합 개념이 정립되면, 하한 및 상한 근사는 유일하게 된다. 그렇지만, 불완전 의사결정표의 경우에 특징 관계와 개념이 단집합, 부분집합 및 개념 근사로 불리는 3가지 상이한 하한 및 상한 근사를 가져올 수 있다. 단집합 하한 및 상한 근사에 관한 여러 연구들이 있지만, 이러한 단집합 근사는 데이터마이닝에 적용될 수 없다[5]. 반면에 개념 하한 및 상한 근사로부터 생성된 규칙들은 아주 유의한 결과들을 나타낸다. 그렇지만, 누락된 속성 값들이 여러 개 존재하는 경우에 상한 및 하한 근사 집합은 다양한 형태를 가지게 되어 생성된 규칙들이 유의하지 않을 수 있다.

본 연구의 2장에서는 의사결정 규칙들의 유도를 위한 트리 기반 분류 알고리즘 및 러프집합의 주요 개념들을 소개하고, 누락된 속성 값들의 처리에 관련된 보편적인 방식들에 관해 논의한다. 3장에서는 누락된 속성 값들이 여러 개 존재하는 경우에 따른 불완전 정보테이블을 완전 정보테이블로 전환시키기 위한 하나의 접근방식을 나타낸다. 접근방식의 기본적인 개념은 정보테이블이 갖는 전체 정보를 속성에 의한 정보, 객체들에 의한 정보 및 속성 및 객체 정보에 의해 설명될 수 없는 오차 부분으로 분할하여 오차에 기인된 정보가 최소가 되는 방식으로의 누락된 속성 값의 대체를 나타낸다. 이는 정보 엔트로피 관점에서 정보 이득을 높이는 방식과 같다. 4장에서는 러프집합 분석과 관련하여 비교적 우수한 프로그램이라 할 수 있는 ROSE(: Rough Set Data Explorer) 프로그램과의 비교 실험을 나타낸다.

## II. 결정 트리 알고리즘과 러프 집합

### 2.1 결정 트리 알고리즘

결정 트리 분석을 위한 대표적인 알고리즘은 ID3이고, ID3의 단점들을 보완한 C4.5 및 C5.0 알고리즘이 사용되고 있으며, 이외에 CART(: Classification and regression trees) 및 CHAID 알고리즘이 적용될 수 있다. 참고로 ID3, C4.5, C5.0은 인공지능 및 기계학습 분야에서 발전된 것이고, CART와 CHAID는 통계 분야에서 개발된 것들이다. 인공지능 계열의 결정 트리 알고리즘들은 엔트로피 개념이 주축이지만, 통계 기반 결정 트리 방법들은 통계적 유의성 검정을 기반으로 하고 있다. 그렇지만, 기본적인 트리 생성 방식은 유사하며, 트리 확장, 즉 가지 분리 방식에서 다소간 차이를 보이고 있다. 문제는 결정 트리 구축을 위한 주요 속성들이 누락되는 경우이다. 통계 기반 방식들인 CART와 CHAID의 경우에 통계적 유의성 검정을 바탕으로 하기 때문에 대체적으로 누락 값이 적은 경우 통계적 추리를 통한 대체를 통해 비교적 원활하게 결정 트리를 구축할 수 있다. 그렇지만, 인공지능 계열의 결정 트리 알고리즘은 엔트로피 개념을 사용하기 때문에 주요 속성 값들이 누락된 경우에 결정 트리가 적용 알고리즘에 따라 달라지는 경향이 자주 나타난다[2].

ID3는 Quinlan에 의해 제안된 대표적인 의사결정트리 기반 분류 알고리즘이다. 이후에 개발된 다양한 의사결정트리 기반의 분류 알고리즘인 C4.5, CART, CHAID 등도 기본적으로 ID3 알고리즘을 바탕으로 하고 있다. ID3 알고리즘과 관련하여 누락된 속성 값들의 처리를 위한 접근방식으로는 Heuristic 기법, 엔트로피, 정보 이득의 개념이 고려될 수 있다. 엔트로피는 주어진 데이터 집합의 혼잡도를 의미한다. 즉, 주어진 데이터 집합에 레코드들이 서로 다른 종류들이 많이 섞여있으면 엔트로피가 높고, 같은 종류의 레코드들이 많이 있으면 엔트로피가 낮다. 엔트로피 값은 0에서 1사이의 값을 갖는데, 가장 혼잡도가 높은 상태의 값은 1, 하나의 클래스로만 구성된 상태의 값이 0이다. ID3 결정트리 알고리즘은 엔트로피가 높은

상태에서 낮은 상태가 되도록 데이터를 특정 조건에서 찾아 트리 형태로 구분해 나간다.

데이터 집합 $S$에 관한 엔트로피 $E(S)$은 다음과 같이 계산된다.

$$E(S) = -\sum_{i=1}^{m} p_i \log_2(p_i),\ p_i = f(C_i, S)/|S| \quad (1)$$

여기서 $p_i$은 $i$번째 클래스 값에 대해 해당 데이터 집합이 차지하는 비율이고, $S$은 주어진 데이터 집합, $C = \{C_1, C_2, ..., C_m\}$ 은 속성 값들의 집합, $f(C_i, S)$은 $S$에서 속성 $C_i$에 속하는 개수, $|S|$은 데이터 개수이다. 따라서 속성 값들의 일부가 누락된 경우에 어떤 값으로 대체하느냐에 따라 엔트로피가 다르게 나타나는 문제점이 존재한다.

정보 이득은 어떤 속성을 선택함으로 인해서 데이터를 더 잘 구분하게 되는 것을 의미한다. 속성 $A$에 관한 정보이득 $Gain(A)$는 속성 $A$를 선택했을 때의 정보이득 양으로 속성 $A$를 선택한 후에 $m$개의 하위 노드로 나누어진 것에 대한 전체적인 엔트로피를 빼는 방식으로 구할 수 있다. 차이가 클수록 정보 이득이 큰 것이고 해당 속성 $A$가 변별력이 좋다는 것을 의미한다.

$$Gain(A) = I(s_1, s_2, ..., s_m) - E(A) \quad (2)$$

여기서 $I(s_1, s_2, ..., s_m)$ 은 상위 노드의 엔트로피이고, $E(A)$은 $A$라는 속성을 선택했을 때 하위 작은 $m$개 노드를 나누어지는 경우에 하위 각 노드의 엔트로피를 계산한 후 노드에 속한 개수를 가중치로 하여 엔트로피를 평균한 값이다. 정보 이득 방식 또한 엔트로피 계산이 요구된다는 점에서 누락된 속성 값들이 존재하는 경우에 특정 값으로의 대체는 변별력이 높은 속성의 정보 이득이 삭감될 수 있는 문제점을 가지고 있다.

C4.5 알고리즘은 Quinlan에 의해 제안된 것으로 ID3와 크게 다르지 않으며, ID3의 몇 가지 단점들을 보완한 형태이다. ID3는 범주 형 속성에 대해서만 트리를 생성하는 방법이기 때문에, 수치형 속성은 모델 생성에 활용할 수 없는 한계가 있다. C4.5에서는 수치형 속성까지 사용하는 방법을 포함한다. 특히, C4.5는 무의미한 속성을 제외하기 때문에(모든 레코드를 잘게 분할하는 속성이 선택되지 않도록 함), 무의미한 속성들에 대한 누락 값은 문제가 되지 않는다. 그렇지만, 연속형 속성에 대해 같은 속성에 대해서도 계속적인 분할이 발생할 수 있기 때문에 연속형 속성 값이 누락된 경우에 문제가 될 수 있다. 주된 이유는 하나의 연속형 속성에 대한 기준이 한꺼번에 나타나지 않고, 여러 경로에 걸쳐서 분산되어 나타나기 때문이다. 이러한 문제점 등으로 인해 연속형 데이터에 대해서는 다원 분할보다는 이진 분할 방식을 사용한다. 이진 분할을 위해서는 하나의 분할 점을 찾아야 한다. 가능한 모든 값에 대하여 분할 점을 설정하고 정보이득 값을 계산해 보는 것이 바람직하지만, 장시간이 소요된다.

CART 알고리즘은 ID3와 접근 방식은 동일하지만, 차이점은 속성 선택을 위한 기준으로 엔트로피 변화가 아닌 엔트로피 행렬을 사용한다는 점이다. 따라서 CART 알고리즘의 장점은 후보 트리를 여러 개 생성하고 그 중에서 최적 트리를 찾아내는데, Gini Index 또는 분산의 감소량을 사용하여 트리 가지를 나눈다. 범주 형 변수에 대해서는 지니 지수를 사용하고, 연속형 변수에 대해서는 분산의 감소량을 사용한다. 따라서 누락된 속성 값이 연속형 변수인 경우에 분산의 감소량은 크게 영향을 받지 않기 때문에 트리 분할이 문제가 되지 않지만, 범주형인 경우에는 Gini Index 값을 가장 많이 감소시켜 주는 변수가 영향을 가장 많이 끼치는 변수가 되고, 이 변수를 기준으로 결정트리 가지가 만들어지기 때문에, 누락 값은 문제가 될 수 있다.

CHAID(Chi-square Automatic Interaction Detection) 알고리즘은 범주 형 속성에 관해서는 카이제곱 검정, 수치형 속성에 관해서는 F-검정을 이용하여 다원 분할을 수행한다. 초기 CHAID는 원래 변수들 간의 통계적 관계를 찾는 것이 주목적이었다. 변수들 간의 통계적인 관계는 다시 의사결정트리를 통해 표현될 수 있으므로, 이 방법은 분류 기법으로 적용될 수 있다. 결정 트리 구축 방식이 CART와 흡사하나 데이터를 분할하는 방식에 차이가 있다. 최적의 분할 혹은 최적 조건 변수를 선택하는데 있어 엔트로피나 지니 매트릭스 대신 통계의 카이제곱 검정을 사용한다는 점에서 통계적 추리에 의한 누락 값 대체는 크게 문제될 것이 없다.

### 2.2 러프 집합

러프집합은 초기 집합을 보통 논리의 상한 및 하한 근사 쌍의 집합으로 근사시키는 하나의 체계적인 방법론에 속한다. 표준 러프집합 이론의 측면에서 하한 및 상한 근사는 보통집합들이라 할 수 있지만, 다른 관점에서 보면 퍼지집합일 수도 있다. 따라서, 특정 객체들의 속성 값이 누락되었을 경우에 러프집합의 형태 및 관련 성질들이 매우 불안정해질 수 있다. 러프집합과 관련하여 흥미로운 문제는 동치류 구조로 표현된 지식에 있어서 어떠한 속성들이 정보 시스템에서 보다 중요한 것이지를 발견하는 것이다. 이러한 문제를 떠나 종종 데이터베이스 내에 지식을 완전하게 특징짓는 속성들의 부분집합을 찾는 것이 쉽지는 않다. 이러한 속성 집합을 Reduct라 한다. 하나의 Reduct는 다음과 같은 속성들의 부분집합 $RED \subseteq P$ 이다.

$[x]_{RED} = [x]_P$: 축소된 속성 집합 $RED$에 의해 유도된 동치류가 전체 속성집합 $P$에 의해 유도된 동치류 구조와 같다.

어떠한 속성 $a \in RED$ 에 관해 $[x]_{(RED-\{a\})} \neq [x]_P$ 라는 의미에서 속성집합 $RED$는 *minimal*이다. 이는 동치류 $[x]_P$의 변경이 없이는 $RED$로부터 어떠한 속성도 제거시킬 수 없음을 의미한다.

Reduct는 특징들에 관해 하나의 범주 형 구조를 나타내기 위한 하나의 충분조건의 집합이다. 하나의 Reduct에 의해 투영된 정보 시스템은 전체 속성 집합에 의해 표현된 같은 동치류 구조를 가지게 된다. 따라서 Reduct는 하나의 대표적인 속성집합으로 집합 원소들 중에 어떤 것이 제거되는 경우에 동치류 구조가 붕괴된다. 문제는 정보 시스템의 Reduct가 유일하지 않다는 점이다. 정보 시스템 내에 표현된 동치류 구조(말하자면, 지식)를 보존하는 많은 속성들의 부분집합이 존재할 수 있다. 모든 Reduct에 공통인 속성 집합은 Core이다. Core는 모든 적절한 Reduct가 보유하고 있는 속성 집합으로 정보 시스템에서 결코 제거될 수 없는 속성들이다. Core는 필요조건의 속성집합으로 간주할 수 있는데, 범주 형 구조를 대표하는 것이다. 즉, Core는 정보시스템의 독립적인 속성이 되는 것이다. Core가 공집합이 될 수도 있는데, 이는 독립적인 속성이 없음을 의미한다[1],[3].

데이터베이스 분석 혹은 데이터 획득 시스템의 가장 중요한 측면 가운데 하나는 속성 종속의 발견이다. 즉, 어떤 변수가 다른 변수들과 강한 관련성을 갖느냐 하는 것이다. 일반적으로 심층적인 조사를 요구하는 강한 관계들이 존재하며, 이러한 것들이 모델 예측에 중요하다. 러프집합에서 종속의 개념은 간단히 정의할 수 있는데, 누락된 속성 값들이 존재하는 경우에 종속 정도의 측정이 유일하지 않다는 점이다. 2개의 배타적인 속성 집합 $P$와 $Q$을 취해서 그들 간에 종속 정도를 측정한다. 각 속성집합은 $P$에 의해 유도된 동치류 $[x]_P$와 $Q$에 의해 유도된 동치류 $[x]_Q$을 가진다. 속성 집합 $Q$에 의해 유도된 하나의 동치류 $Q_i$에 관해, $[x]_Q = \{Q_1, ..., Q_N\}$ 라 하면, 속성집합 $P$상에서 속성 집합 $Q$의 종속성 $\gamma_P(Q)$ 은 다음과 같이 측정될 수 있다.

$$\gamma_P(Q) = \frac{\sum_{i=1}^{N} |\underline{P}Q_i|}{|U|} \leq 1 \tag{3}$$

즉, $[x]_Q$에서 각 동치류 $Q_i$에 관해 $P$에 포함된 속성들에 의한 하한 근사의 크기($\underline{P}Q_i$)를 더한다. 이러한 근사는 속성집합 $P$상에서 목표 집합 $Q_i$에 속하는 것으로 명확히 식별할 수 있는 객체들의 개수가 된다.

$\gamma_P(Q)$ 은 $Q$에 포함된 속성 값들의 결정을 위해 $P$에 포함된 속성 값들을 알고자 하는 경우에 정보시스템에서 이러한 객체들의 비율로 해석될 수 있다. 종속성을 고려하기 위한 또 다른 직관적인 방식은 $Q$에 의해 유도된 분할을 목표 클래스 $C$로 간주하고, $P$을 목표 클래스 $C$의 재구축을 위한 속성 집합으로 사용하는 것이다. 만일, $P$에 의해 $C$가 완전하게 구축된다면, $Q$은 전반적으로 $P$에 종속되는 것이다. $P$에 의한 결과가 빈약해서 $C$에 관한 하나의 무작위 구축이 된다면, $Q$은 $P$에 전혀 종속적인 것이 아니게 된다. 종속성의 이러한 측도는 속성 집합 $P$상에서 속성 집합 $Q$의 함수적인 종속 정도를 나타낸다[8].

러프집합 이론은 불완전한 데이터 집합에서 이러한 종속성을 바탕으로 하는 규칙 유도에 유용하기 때문에, 누락된 속성에 대한 처리가 매우 중요하다. 러프집합과 관련하여 누락된 속성 값들은 3가지 유형으로 구분할 수 있다. 하나는 lost 값으로 이것은 기록되어 있으나 현재 이용 가능하지 않은 것이고, 다른 하나는

속성-개념 값으로 같은 개념에 속하는 어떠한 속성 값으로 대체될 수 있는 것이다. 마지막으로 원래 값이 부적절한 “do not care” 조건이다. 하나의 개념 혹은 클래스는 같은 방식으로 분류된 모든 객체 집합이다. 누락된 속성 값들을 갖는 두 가지 특수한 형태의 데이터 집합이 논의되어왔다[7].

## III. 정보 변동 기반 누락 속성 처리와 관련 규칙

### 3.1 정보 테이블 재구성과 변동 추정

러프집합 이론 기반의 의사결정 규칙 도출에 있어서 문제가 되는 것은 누락된 속성 값들이 여러 개가 존재하는 경우이다. 앞 장에서 논의된 것처럼 특정 객체들의 속성 값이 누락되었을 경우에 러프집합의 형태 및 관련 성질들이 매우 불안정해진다. 누락된 속성 값에 관해 속성-개념 값 해석은 누락된 속성 값을 갖는 객체가 속할 가능성이 높은 개념의 속성 값으로 대체시키는 것이 일반적인 접근방식이다. 본 장에서는 누락된 속성 값의 추정과 관련하여 해당 속성 정보를 중심으로 하는 추리보다는 전체 속성 정보의 관점에서 누락된 속성 값의 정보이득을 줄이는 방식으로의 전처리 방식을 나타내고, 이를 통한 의사결정규칙들의 유도를 나타낸다.

조건 속성 값들이 누락된 정보 테이블은 하나의 불완전 의사결정표가 된다. 러프집합과 관련하여 누락된 속성 값들은 보통 “?” 혹은 “*”로 표시된다. 여기서 “?”의 누락된 값은 원래 속성 값이 알려져 있었으나 어떠한 알 수 없는 이유로 인해 삭제된 경우이고, “*”에 의한 값은 부적절한 이유로 인해 원래 값들이 기록되지 않은 경우이다[3]. 본 논문에서는 각 객체에 관해 적어도 하나의 속성 값은 존재하는 것으로 가정한다. 즉, 일부 조건 속성 값들은 누락되어 있지만, 모든 의사결정 값들은 존재하는 것으로 한다. 문제는 누락된 속성 값들이 많거나, 보편적인 속성 값들의 빈도가 거의 균등한 경우이다. 누락된 속성 값이 희소하거나, 속성이 갖는 범주들의 빈도가 확연하게 차이가 있을 경우에는 가장 보편적인 속성 값들로 누락된 속성 값들을 대체하거나 또는 자주 나타나는 속성 값으로 모든 누락된 속성 값들을 대체하는 방식이 단순하면서도 다른 복잡한 방법보다 크게 성능이 떨어지지 않는다. 그렇지만, 누락된 속성 값이 많거나 또는 속성 값들의 히스토그램이 균등한 경우에는 이러한 대체 방식은 적절하지 않고, 정보테이블에 의문이 생겨 얻어진 데이터 자체도 신뢰할 수 없는 것이 된다.

$n$개의 속성 $P_1, P_2, ..., P_n$와 $l$개의 객체 $O_1, O_2, ..., O_l$로 구성되는 하나의 정보테이블을 고려하자. 속성 $P_j$에 관한 객체 $O_i$의 속성 값을 $p_{ij}$, 속성 값들의 전체 평균을 $\overline{\overline{p}}$라 하면, $p_{ij}$와 $\overline{\overline{p}}$간에 차이는 다음과 같이 나눌 수 있다.

$$\left(p_{ij}-\overline{\overline{p}}\right)=\left(\overline{p_{i.}}-\overline{\overline{p}}\right)+\left(\overline{p_{.j}}-\overline{\overline{p}}\right)+\left(p_{ij}-\overline{p_{i.}}-\overline{p_{.j}}+\overline{\overline{p}}\right) \quad (4)$$

위 식의 양변을 제곱하여 정리하면, 다음과 같은 식이 성립한다.

$$\sum_{i=1}^{l}\sum_{j=1}^{n}\left(p_{ij}-\overline{\overline{p}}\right)^2=\sum_i\sum_j\left(\overline{p_{i.}}-\overline{\overline{p}}\right)^2+ \quad (5)$$
$$\sum_i\sum_j\left(\overline{p_{.j}}-\overline{\overline{p}}\right)^2+\sum_i\sum_j\left(p_{ij}-\overline{p_{i.}}-\overline{p_{.j}}+\overline{\overline{p}}\right)^2$$

위 식에서 $V_T=\sum_i\sum_j\left(p_{ij}-\overline{\overline{p}}\right)^2$을 전체 변동,

우측에서 $V_O=\sum_i\sum_j\left(\overline{p_{i.}}-\overline{\overline{p}}\right)^2$을 객체 변동,

$V_A=\sum_i\sum_j\left(\overline{p_{.j}}-\overline{\overline{p}}\right)^2$을 속성 변동,

$V_E=\sum_i\sum_j\left(p_{ij}-\overline{p_{i.}}-\overline{p_{.j}}+\overline{\overline{p}}\right)^2$을 오차 변동

이라 하자. 즉 정보의 전체 변화는 속성 및 객체에 의한 변화로 구분할 수 있다.

표 1. 재구성된 정보테이블
Table 1. Reconstructed information table

| | $A_1$ | $A_2$ | … | $A_j$ | … | $A_n$ | 계 |
|---|---|---|---|---|---|---|---|
| $B_1$ | $p_{11}$ | $p_{12}$ | … | $p_{1j}$ | … | $p_{1n}$ | $S_{1.}$ |
| $B_2$ | $p_{21}$ | $p_{22}$ | … | $p_{2j}$ | … | $p_{2n}$ | $S_{2.}$ |
| ⋮ | ⋮ | ⋮ | | ⋮ | | ⋮ | |
| $B_i$ | $p_{i1}$ | $p_{i2}$ | … | $p_{ij}^m$ | … | $p_{in}$ | $S_{i.}^m$ |
| ⋮ | ⋮ | ⋮ | | ⋮ | | ⋮ | |
| $B_l$ | $p_{l1}$ | $p_{l2}$ | … | $p_{lj}$ | … | $p_{ln}$ | $S_{q.}$ |
| 계 | $S_{.1}$ | $S_{.2}$ | | $S_{.j}^m$ | | $S_{.n}$ | $S^m$ |

표 1. 과 같이 재구성된 정보테이블을 고려하자. 표에서 $p_{ij}^m$은 속성 $j$에 관해 $i$번째 행의 값이 누락된 경우이다. 누락된 값이 많은 경우에 $p_{ij}^m$을 기준으로 누락된 값이 없는 행과 열들로 축소된 행렬을 구축한다. 여기서 $S$들은 각각 행 및 열 합을 나타내고, $S_{i.}^m$은 누락된 값을 제외한 행 합, $S_{.j}^m$은 누락된 값을 제외한 열 합, $S$은 누락된 속성 값을 제외한 전체 합을 나타낸다. 누락된 값 $p_{ij}^m$의 추정을 위해 먼저 오차 변동 $V_E = V_T - V_O - V_B$ 을 구한다. 다음으로 $V_E$을 최소로 하는 $p_{ij}^m$을 구하기 위해 $V_E$을 $p_{ij}^m$로 미분하여 0으로 놓으면 다음과 같다.

$$\frac{dV_E}{dp_{ij}^m} = 2p_{ij}^m - \frac{2}{l}\left(p_{ij}^m + S_{.j}^m\right) - \frac{2}{n}\left(p_{ij}^m + S_{i.}^m\right) + \frac{2}{nl}\left(p_{ij}^m + S^m\right) = 0 \quad (6)$$

$p_{ij}^m$에 관해 풀면 다음과 같은 추정 값을 얻을 수 있다.

$$p_{ij}^m = \frac{nS_{.j}^m + lS_{i.}^m - S^m}{(l-1)(n-1)} \quad (7)$$

여기서 추정된 $p_{ij}^m$의 값은 실수 값이기 때문에 범주 형의 값으로 전환시키기 위해 $p_{ij}^m$의 반올림을 취한 정수 값을 사용할 수 있다.

### 3.2 결정 규칙의 타당성 검토

정보 변동을 기반으로 하는 누락된 속성 값들의 대체를 통해 하나의 완전 의사결정표가 얻어지면, 다음과 같은 단계로 의사결정 규칙들의 타당성 검토를 진행할 수 있다.

[단계 1] 오차 변동을 최소화시키는 속성 값을 추정하여, 즉 누락된 속성에 관해 오차 변동이 최소가 되는 값으로 대체하는 방식으로 불완전 의사결정표를 완전 의사결정표로 변환시킨다.

[단계 2] 완전 정보 테이블의 Reduct와 Core를 발견하고, 이것들과 러프집합 분석 관련 프로그램과 결과를 비교한다.

[단계 3] 완전 정보 테이블의 Reduct를 바탕으로 규칙들을 생성한다.

[단계 4] 러프집합 관련 프로그램에 의한 것과 규칙들을 비교한다.

Reduct와 Core를 동시에 찾기 위해서는 일치 및 불일치의 개념을 이해할 필요가 있다. 하나의 정보테이블이 일치성을 갖기 위해서는 조건 속성의 모든 같은 값들이 같은 결정 속성을 유도해야 한다. 이에 반하여, 조건 속성의 같은 값들이 상이한 결정 속성을 갖는 경우를 불일치성이라 한다. 정보테이블로부터 속성들 가운데 어느 하나를 제거시킨 후에 나머지 정보테이블로부터 일치성 문제를 검토할 수 있다. 만일 나머지 정보테이블이 일치성을 보인다면 하나의 Reduct 집합을 형성한다고 할 수 있다. 만일 불일치성을 나타낸다면, 표에서 제거된 속성은 Core 속성이 된다. 불완전 정보테이블로부터 Reduct 및 Core 결정은 매우 중요하지만, 누락된 값들이 많은 경우에 기존 방법에 의한 누락된 값의 대체는 여러 개의 상이한 Reduct 집합 및 Core 집합을 생성하게 된다. 따라서 누락된 값들의 결정에 있어서 속성들에 의한 변동이 순수 오차 변동에 의한 것보다 크게 되는 추정 방법이 중요하다. 이러한 방식에 의한 Reduct 결정은 다른 대체 방식에 비해 다소간 시간 복잡도가 증가하지만, 유일한 규칙들을 생성할 수 있다.

## Ⅳ. 구현

이 장에서는 본 연구에서 제안된 정보 변동에 비해 속성 변동이 가장 큰 값으로 누락된 속성 값을 대체하는 방식에 가장 근접된 결과를 보이는 ROSE(: Rough Sets Data Explorer) 프로그램의 활용과 이를 통한 의사결정 규칙 결정에 관한 비교를 나타낸다. ROSE는 러프집합 이론 및 규칙 발견 기법의 기본적인 사항들을 구현하는 프로그램이다. 대부분의 연산은 Pawlak에 의해 소개된 러프집합 개념들을 바탕으로 하고 있지만, Ziarko에 의해 제안된 변수 적정 러프집합 모델을 적용하고 있다. 이러한 변수 적정 모델 구축의 개념은 통계적으로 관련 변수 변동이 오차 변동에 비해 큰 것을 선택한다는 점에서 정보테이블의 누락된 속성 값 대체 방식에서 본 연구의 오차 변동 최소화 개념과 일치한다고 할 수 있다.

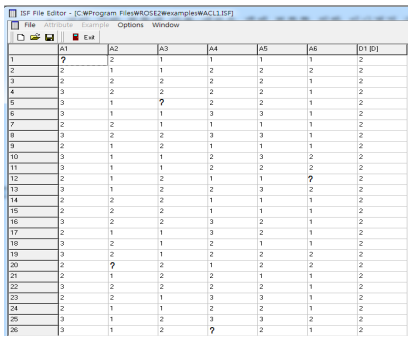

그림 1. 누락된 속성 값
Fig. 1 Missing attribute value

그림 1.은 ROSE의 Example 폴더에서 ACL1.ISF의 정보테이블을 로드시킨 경우이다. 본 연구에서 제안된 정보 변동의 최소화에 의한 누락된 값의 처리를 위해 속성 A1의 첫 번째 사례, 속성 A2의 20번째 사례, 속성 A3의 5번째 사례, 속성 A4의 26번째 사례, 속성 A6의 12번째 사례가 누락된 경우를 가정한다. 이렇게 설정한 이유는 해당된 사례의 속성 값이 누락된 경우에 '자주 나타나는 속성 값' 혹은 '가장 보편적인 속성 값'으로 대체하는 경우에 모두 2가 될 가능성이 높기 때문이다. 먼저 속성 A1의 누락된 값을 추정하기 위해 5개 사례를 제외한 135개 사례를 활용한 추정 결과는 2.5612로 가까운 범주 값은 3이 된다. 두 번째로 속성 A2의 누락된 값은 1.4937로 추정 값은 1이고, 세 번째 A3의 5번째 사례의 추정 값은 1.4918로 1로 추정되었고, 네 번째 A4의 26번째 사례의 추정 값은 2.0285로 2로 추정되고, 다섯 번째 A6의 12번째 사례의 추정 값은 1.1313으로 1로 추정되었다.

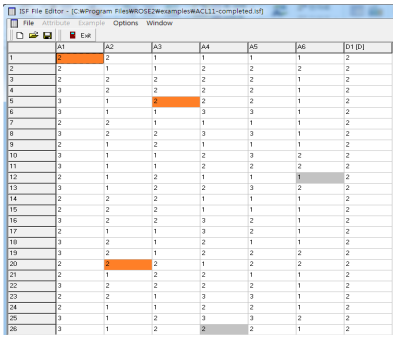

그림 2. ROSE에 의한 누락된 값 처리
Fig. 2 Missing value estimation by ROSE

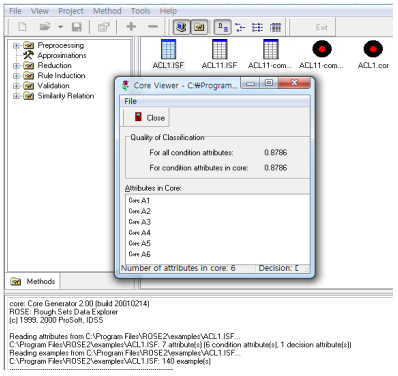

그림 3. 제안된 방법에 의한 전처리와 Core
Fig. 3 Preprocessing by suggested method and Core

추정 결과는 모두 원천 데이터와 일치되어 누락된 값의 추정에 있어서 제안된 방식에 의한 전처리가 타당함을 나타냈다. 참고로 ROSE 프로그램에 의한 추정 값은 그림 2.와 같다. 그림에서 주황색 부분은 원천 데이터와 다르게 나타난 부분이고, 회색 부분은 일치한 경우로 전체적으로 40%의 적중률을 나타낸다. 이는 가장 보편적인 속성 값으로 대체한 결과와 같은 것이 된다. 물론 본 연구의 추정 결과를 절단 값으로 대체한 경우에는 그림 2.의 경우와 같은 결과가 얻어진다. 그림 3.은 본 연구에 의한 전처리를 통한 Core를 계산한 결과이고, 그림 4.는 ROSE에 의한 전처리와 Core를 구한 것이다. 분류 성능은 본 연구의 전처리에 의한 것이 (0.8786) ROSE2에 의한 것보다 (0.8714) 나은 것으로 판정되었다.

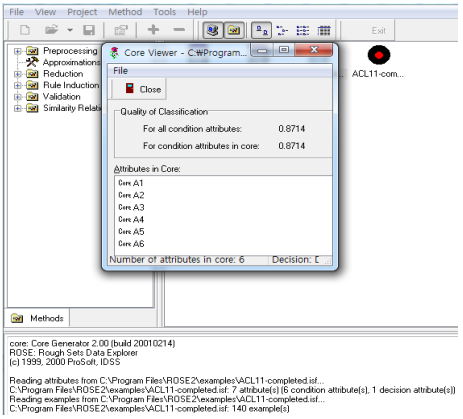

그림 4. ROSE에 의한 전처리와 Core
Fig. 4 Preprocessing and Core by ROSE

## Ⅴ. 결론

본 연구에서는 누락된 속성 값들을 갖는 객체들을 제외한 전체 정보 대비 속성 정보의 변동을 고려하는 추정 방법을 도입하였다. 러프집합 분석과 관련하여 비교적 성능이 우수한 프로그램이라 할 수 있는 ROSE2 프로그램에 의한 전처리 방법과 비교한 결과 Rose 프로그램 추정에 의한 결과는 원천 데이터와 달라질 수 있지만, 제안된 방식에 의한 추정은 원천 데이터의 값과 대부분 일치함을 보여 보다 견고한 누락 속성 데이터의 추정 방법이라 할 수 있다. 다른 무엇보다도 중요한 점은 하나의 누락된 속성 값이 잘못 대체되는 경우에 Reduct와 Core의 변동은 없지만, 의사결정 규칙이 달라질 수 있으나, 제안된 방법의 장점은 원천 데이터에 의해 생성되는 규칙들을 보존하는 보다 견고한 특성을 보였다. 제안된 방법의 보완점은 거대 정보테이블에 적용하는 경우에 기존 대체 방법에 비해 소요시간이 길어진다는 점이고, 입력 에러에 따른 잘못된 값들을 검토할 수 없기 때문에 이에 관한 보완 연구가 필요하다.

## References


[1] J. Bazan, M. Szczuka, and A. Wojna, "On the evolution of rough set exploration system," *Proc. of the Rough Sets and Current Trends in Computing,* Uppsala, Sweden, June, 2004, pp. 592 - 601.

[2] G. Claeskens and N. L. Hjort, *Model Selection and Model averaging.* England Cambridge: Cambridge University Press. 2004.

[3] V. Dubois and M. Quafafou, "Concept learning with approximation: Rough version spaces," *Rough Sets and Current Trends in Computing: Proc. of the 3-rd Int. Conf., Rough Sets and Current Trends in Computing,* Malvern, PA, Oct. 2002, pp. 239 - 246.

[4] J. W. Grzymala-Busse, "Rough set strategies to data with missing attribute values," *Proc. of the Workshop on Foundations and New Directions of Data Mining, the 3-rd Int. Conf. on Data Mining,* Melbourne, FL, USA, Nov 19-22. 2003, pp. 56-63.

[5] J. W. Grzymala-Busse and M. Hu, "A comparison of several approaches to missing attribute values in data mining," *Proc. of the Second Int. Conf. on Rough Sets and Current Trends in Computing Rough Sets and Current Trends in Computing,* Banff, Canada, Oct. 2000, pp. 340-347.

[6] Z. Pawlak and A. Skowron, "Rudiments of rough sets," *Information Sciences*, 1 January 2007. vol. 177 no. 1, pp. 3-27.

[7] J. Stefanowski and A. Tsoukias, " Incomplete information tables and rough classification," *Computational Intelligence,* vol. 17 no. 3, August 2001. pp. 545-566.

[8] J. T. Yao and Y. Y. Yao, "Induction of classification rules by granular computing," *Proc. of the Third Int. Conf. on Rough Sets and Current Trends in Computing (TSCTC'02).* London, UK, Sept. 2002, pp. 331 - 338.

[9] D. Cha, K. Ban and E. Kim,"Schema Mapping Method using Frequent Pattern Mining," *J. of the Korea Institute of Electronic Communication Sciences,* vol. 5, no. 1, 2013, pp.96-99.

[10] S. Cho, "A Fuzzy-based Fusion Wireless Localization Method"*J. of the Korea Institute of Electronic Communication Sciences,* vol. 10, no.4, 2015, pp.508-510.


## 저자 소개


**이영천(young-cheon Lee)**

1978년 전남대학교 수학과 졸업(이학사)
1981년 조선대학교 대학원 수학과 졸업(이학석사)
2012년 전남대학교 대학원 컴퓨터공학과 졸업(공학석사)
1991년 조선대학교 대학원 수학과 졸업(이학박사)
1993년~2003년 호남대학교 수학과 교수
2004년~현재 호남대학교 컴퓨터공학과 교수
※ 관심분야 응용수학(암호이론, 최적화 기법)