韓國컴퓨터情報學會 論文誌
第14卷 第6號, 2009. 6.

2009-14-6-5-1

데이터마이닝의 자동 데이터 규칙 추출 방법론 개발 : 계층적 클러스터링 알고리듬과 러프 셋 이론을 중심으로

오 승 준 *, 박 찬 웅**

Development of Automatic Rule Extraction Method in Data Mining : An Approach based on Hierarchical Clustering Algorithm and Rough Set Theory

Seung-Joon Oh *, Chan-Woong Park **

요 약

데이터 마이닝은 대용량의 데이터 셋을 분석하기 위하여 새로운 이론, 기법, 분석 툴을 제공하는 전산 지능분야의 새로운 영역중 하나이다. 데이터 마이닝의 주요 기법으로는 연관규칙 탐사, 분류, 클러스터링 등이 있다. 그러나 이들 기법을 기존 연구 방법들처럼 개별적으로 사용하는 것보다는 통합화하여 규칙들을 자동적으로 발견해내는 방법론이 필요하다. 이런 데이터 규칙 추출 방법론은 대량의 데이터들을 분석하여 성공적인 의사결정을 내리는데 도움을 줄 수 있기에 많은 분야에 이용될 수 있다. 본 논문에서는 계층적 클러스터링 알고리듬과 러프셋 이론을 이용하여 대량의 데이터로부터 의미 있는 규칙들을 발견해 내는 자동적인 규칙 추출 방법론을 제안한다. 또한 UCI KDD 아카이브에 포함되어 있는 데이터 셋을 이용하여 제안하는 방법에 대하여 실험을 수행하였으며, 실제 생성된 규칙들을 예시하였다. 이들 자동 생성된 규칙들은 효율적인 의사결정에 도움을 준다.

Abstract

Data mining is an emerging area of computational intelligence that offers new theories, techniques, and tools for analysis of large data sets. The major techniques used in data mining are mining association rules, classification and clustering. Since these techniques are used individually, it is necessary to develop the methodology for rule extraction using a process of integrating these techniques. Rule extraction techniques assist humans in analyzing of large data sets and to turn the meaningful information contained in the data sets into successful decision making. This paper proposes an autonomous method of rule extraction using clustering and rough set theory. The experiments are carried out on data sets of UCI KDD archive and present decision rules from the proposed method. These rules can be successfully used for making decisions.

▸ Keyword : 규칙추출(rule extraction), 러프셋(rough set), 리덕트(reduct)

• 제1저자 : 오승준
• 투고일 : 2009. 05. 14, 심사일 : 2009. 05. 14, 게재확정일 : 2009. 06. 11.
* 경기공업대학 산업경영과 교수 ** 경원대학교 산업정보시스템공학과 교수

Ⅰ. 서 론

데이터 마이닝이란 대용량의 데이터 셋을 분석하기 위하여 새로운 이론, 기법, 분석 툴을 제공하는 전산 지능분야의 새로운 영역중 하나이며, 기계학습, 클러스터 분석, 회귀 분석, 뉴럴 네트워크 등과 관련이 있는 분야이다. 최근에는 대량의 데이터들이 디지털 형태로 제공됨에 따라 이 분야에 대한 연구가 활발해 지고 있다.

최근 정보 산업 분야에서 데이터 마이닝이 주목받고 있는데 그 주된 이유는 데이터의 양적 팽창과 그러한 데이터를 유용한 정보와 지식으로 바꿔야 하는 필요성 때문이다. 이렇게 얻어진 정보와 지식은 기업경영, 생산운영 그리고 시장분석에서부터 공학설계와 과학탐구에 이르기까지 광범위한 응용 분야에 이용될 수 있다.

데이터 마이닝의 주요 기법으로는 연관규칙 탐사, 분류, 클러스터링 등이 있는데, 이들은 각각 개별적으로 데이터들을 분석하여 연관규칙을 찾아내거나 분류 모형을 만들거나 데이터들을 군집화 하는데 사용된다[1,2,3]. 그러나 이들 기법을 개별적으로 사용하는 것보다는 통합화하여 체계적인 방법으로 의미 있는 규칙들을 발견해 내는 새로운 접근 방법론이 필요하다. 이런 데이터 규칙 추출 방법론은 인간이 대량의 데이터들을 분석하여 이들 정보로부터 성공적인 의사결정을 내리는데 도움을 줄 수 있기에 많은 분야에 이용될 수 있다.

본 논문에서는 계층적 클러스터링 알고리듬을 이용하여 러프셋 이론 기반의 자동 규칙 추출 방법을 제안한다. 이 방법은 대량의 데이터로부터 의미 있는 규칙들을 발견해 내는 체계적인 방법론이다. 여기서 사용되는 데이터들은 범주형 뿐만 아니라 이산형 데이터도 가능하며, 결정 변수(클래스 속성)가 없는 데이터들을 대상으로 한다. 본 논문의 제안 방법은 다음과 같이 크게 네 가지 단계로 구분할 수 있다. 첫 번째는 데이터들을 전처리 하는 과정이다. 결측치나 수치형 데이터에 대한 이산화 작업을 수행하는 과정이다. 두 번째 단계는 데이터들을 클러스터링 하는 단계이다. 데이터들을 계층적 클러스터링 알고리듬을 이용하여 몇 개의 그룹들로 나눈 후, 이 결과를 결정 변수로 사용한다. 세 번째 단계는 두 번째 단계에서 얻어진 데이터들에 러프셋 이론을 적용하여 리덕트들을 찾아내는 것이다. 마지막 단계에서는 전 단계에서 찾아낸 리덕트들을 이용하여 의미 있는 규칙들을 발견해 내는 것이다.

Ⅱ. 관련 연구

1. 클러스터링

데이터 마이닝 기법에서 클러스터링이란 물리적 혹은 추상적 객체들을 서로 비슷한 객체들의 클래스로 그룹화 하는 과정으로, 하나의 클러스터에 속하는 객체들 간에는 서로 다른 클러스터 내의 객체들과는 구분이 되는 유사성을 갖게 된다.[1] 클러스터링 기법들은 통계학(statistics), 패턴인식(pattern recognition), 데이터 분석, 이미지 처리 그리고 시장조사를 포함한 매우 많은 응용분야에서 넓게 사용된다.

모집단이란 m 개의 속성들로 이루어진 n 요소들의 집합이다. 클러스터링의 목표는 적당한 유사도 측정 방법에 의하여 데이터들을 유사한 클러스터들로 그룹화 하는 것이다. 즉, 유사한 데이터들은 동일 클러스터에 할당이 되는 반면에 확실하게 구별이 되는 데이터들은 서로 다른 클러스터들에 할당이 되도록 하는 것이다.

클러스터링 기법들은 크게 계층적(hierarchical) 방법과 분할(partition) 방법으로 나눌 수 있다. 계층적 방법은 주어진 데이터들을 계층적으로 분해해 나가는 방법으로, 어떻게 계층적 분해가 이루어지는가에 따라 통합적이거나 분리적 방법으로 나누어진다. 통합 방법은 처음에 각각의 객체들을 하나의 클러스터로 설정 한 후 이들 쌍간의 거리 (혹은 유사도)를 기반으로 가장 가까운 클러스터(객체)들끼리 합병을 수행한다. 최종적으로 한 클러스터 내에 모든 객체들이 포함될 때까지 위의 과정을 반복한다. 분리 방법은 통합 방법과 반대로 위의 과정을 진행 한다

분할 방법은 어떠한 범주 함수를 최적화 시키는 k개의 분할을 결정해 나가는 방법으로 유클리드안 거리 측정법에 기반을 둔다.

2. 러프 셋 이론

1980년대 초에 Pawlak에 의해 소개된 러프 셋 이론은 어떤 집합에서 확실하게 분류되는 하한 근사 공간(Lower Approximation)과 불확실하게 분류되는 상한 근사 공간(Upper Approximation)을 집합 이론을 통해서 나타낸다[4,5]. 러프 셋 이론의 가장 중요한 장점 중 하나는 부정확하거나 불완전 하고, 애매모호한 성질을 가진 데이터의 분류 분석 문제에 적합한 알고리즘이라는 것이며, 또한 데이터에 대해서 어떠한 사전정보나 부가적인 정보가 필요 없다는 것이다.

동치 관계에 의해 정보 객체 집단은 동치류(equivalence class)로 구분될 수 있으며, 이들 동치류 원소의 집합을 기본 집합이라 하고, 이 기본 집합에 의해 정의되는 집합 공간을

근사(approximation) 공간이라고 한다. 근사 공간상에 하나의 결정에 대한 정보 객체를 분류하는 경우, 동일한 기본 집합 내에 있으면서도 서로 다른 결정을 내는 경우가 발생할 수 있다. 이런 결정상의 불일치(inconsistency)를 나타내고 처리하기 위해서 리프 셋 이론에서는 두 가지 근사를 정의한다.

하나는 결정에 의해 나타내어지는 개념 X에 항상 포함되는 기본 집합으로 정의되는 하한 근사이고, 다른 하나는 개념 X와 일치하는 부분이 하나라도 존재하는 모든 기본집합으로 정의되는 상한 근사이다.

U를 전체 집합이라 하고 R를 U에 대한 동치 관계라 하자. A=(U, R)은 근사 공간이 되며, 하한 근사와 상한 근사는 다음과 같이 표현된다.

$$\underline{R}X = \{x \in U : [x] \subseteq X\},$$
$$\overline{R}X = \{x \in U : [x] \cap X \neq \emptyset\},$$

여기서 [x]는 원소 x를 포함하는 R의 동치류를 나타낸다. X의 러프 셋은 다음과 같이 정의된다.

$$AR(X) = (\underline{R}X, \overline{R}X)$$

또한, 상한 근사에서 하한 근사를 제외시키면 불확실한 개체들만 또 다른 부분 집합으로 표현될 수 있으며, 이를 경계영역이라 부르며, 다음과 같이 표현한다.

$$BNR(X) = \underline{R}X - \overline{R}X$$

전체 집합 U에서 속성 집합 C의 하나의 원소 a가 IND(C) = IND(C-{a})를 만족할 경우, 속성 a는 C에서 불필요(dispensable)하고, 그렇지 않으면 a는 C에서 필요(indispensable)이다. 속성 집합 C에서 임의의 속성을 추출함에 있어서 불필요한 속성집합을 뺀 최소속성집합을 C′라고 할 때 C′ ⊂ A이고 IND(C) = IND(C′)일 경우, C′를 C의 리덕트라고 한다. C′에서 필요인 모든 속성들의 모임, 즉 리덕트들의 교집합을 C의 코어라 한다.

러프 셋 이론을 이용한 연구는 현실세계에 존재하는 불확실한 데이터를 다루는 데 있어 매우 유용하다는 평가를 받고 있기에 적용 분야 또한 대단히 넓다. 예를 들어 인공지능, 인지과학, 의료 데이터 분석, 패턴인식 등과 같은 응용 분야가 있으며, 데이터 마이닝 분야에도 널리 활용 되고 있다. 러프 셋 이론을 분류 기법에 이용한 연구로는 Kim[6]가 있으며, 은행의 파산을 예측하기 위한 연구로는 McKee et. al.[7]가 있으며, 반도체 제조 공정에 이용한 연구로는 Kusiak[8]가 있다. 또한 feature selection에 대한 연구로는 Thangavel[9]가 있으며, 구간 데이터에 대한 연구로는 Asharaf[10]이 있다.

그러나 기존 러프셋 연구들은 대부분 데이터 마이닝 중 분류 기법이나 클러스터링 등 개별 기법에 적용한 연구들로서, 본 연구에서 제안하는 방법처럼 통합적인 방법으로 적용되지 않았다. 자동적인 규칙 추출 방법에 대한 연구로는 Kusiak[8], Kusiak et. al.[11], Sakai[12] 등이 있으나, 이들 모두 본 연구에서 제안하는 방법과 같이 클러스터링을 이용하여 클래스를 부여한 후 규칙들을 생성하는 것이 아니라 클래스가 있는 의사결정 테이블에서 규칙들을 추출하는 방법들에 대해 연구하였다.

III. 4단계 통합 자동 데이터 규칙 추출 방법론

1. 제안하는 방법의 개요

의사결정 테이블은 그림 1과 같은 형태로 표현된다. 열은 n개의 속성들과 결정 변수(클래스 속성)로 구성되며, 행은 m개의 데이터 객체들로 이루어진다.

	속성 1	속성 2	…	속성 n	결정 변수
데이터 객체 1	d11	d12		d1n	d1
데이터 객체 2	d21	d22		d2n	d2
…					
데이터 객체 m	dm1	dm2		dmn	dm

그림 1. 의사결정 테이블
Fig 1. Decision Table

그림 2의 예제 데이터 집합은 5개의 튜플들로 구성되어 있고, 3개의 속성집합 A1, A2, A3 을 가지고 있다. A1속성에는 a, b, c라는 3가지 속성 값을 가지고 있고, A2속성에는 m, f라는 2가지 속성 값을 가지고 있다. 또한 이 데이터 셋에는 결정변수가 주어져 있지 않다.

	A1	A2	A3
1	b	f	z
2	a	m	z
3	c	f	y
4	b	m	z
5	c	f	z

그림 2. 예제 데이터
Fig 2. Example data

본 논문에서는 그림 2의 데이터 셋처럼 결정 변수가 없는 데이터들을 입력 데이터로 사용하여 규칙들을 자동적으로 생성한다.

제안하는 방법은 그림 3과 같은 단계로 이루어진다.

```
Step 1. Data preprocessing
        (missing value, discretization)

Step 2. Hierarchical clustering algorithm
        (class decision)

Step 3. Reduct creation using rough set theory
        (data reduction)

Step 4. Rule extraction using reduct
        (rule generation)
```

그림 3. 제안하는 방법
Fig. 3. The proposed algorithm

제안하는 방법으로 생성된 규칙은 if-then 형태의 규칙으로 다음과 같이 표현된다.

if (속성 i = d_{mi}) and (속성 j = d_{mj})
then 결정변수 = d_m

기존 연구들은 데이터 마이닝의 주요 기법들인 연관규칙 탐사, 분류, 클러스터링 등을 데이터 셋에 개별적으로 적용하여 의미있는 규칙이나 패턴들을 찾아낸다. 즉, 데이터들을 분석하여 연관규칙을 찾아내거나 분류 모형을 만들거나 데이터들을 군집화 한다. 그러나 이들 기법을 개별적으로 사용하는 것보다는 통합화하여 체계적인 방법으로 의미 있는 규칙들을 발견해 내는 새로운 접근 방법론이 필요하다. 이런 데이터 규칙 추출 방법론은 인간이 대량의 데이터들을 분석하여 이들 정보로부터 성공적인 의사결정을 내리는데 도움을 줄 수 있기에 많은 분야에 이용될 수 있다. 즉 여러 단계를 개별적으로 수행하며 규칙들을 찾아내기에는 많은 시간과 시행착오가 필요하므로, 본 연구에서 제시하는 통합적인 방법의 자동 규칙 추출 방법론이 현실 문제의 효율적인 의사결정에 많은 도움을 준다.

2. 데이터 전처리 과정

첫 번째 단계인 데이터 전처리 과정에서는 크게 결측치에 대한 처리와 수치형 데이터들을 범주형 데이터로 이산화 하기 위한 처리 과정을 수행한다. 본 논문에서는 결측치를 처리하기 위해서 결측치가 포함되어 있는 데이터를 제거하는 과정을 수행한다.

데이터 속성들의 값들은 범주형 뿐만 아니라 수치형 값으로도 이루어진다. 따라서 본 논문에서 사용되는 데이터들은 범주형 뿐만 아니라 이산형 값들도 사용될 수 있도록 하기 위하여 이산형 값들을 가진 속성들을 범주형 값으로 변환하기 위한 이산화 처리 과정을 수행한다. 이를 위해 엔트로피를 기반으로 한 Dougherty et. al.[13]가 제안한 알고리듬을 이용하여 이산화 과정을 수행한다.

3. 계층적 클러스터링 알고리듬 적용

본 논문에서는 통합 방법의 계층적 클러스터링 알고리즘을 사용한다. n개의 데이터들을 클러스터링 하는 문제를 생각해 보자. 처음에는 n·(n-1)/2개의 클러스터간 합병을 고려할 수 있는데, 이중에서 합병을 했을 경우 가장 높은 평가함수 값을 주는 두 개의 클러스터를 합병한다. 1번째 합병 후에는 (n-1)·(n-1-1)/2개의 클러스터간 합병을 고려하며, 이중에서 가장 높은 평가함수 값을 주는 두 개의 클러스터를 합병한다. 최종적으로는 주어진 개수의 클러스터가 남을 때까지 위의 과정을 반복한다.

본 논문에서는 평가함수로 식(1)을 사용한다.

$$\text{maximize } Cf = \sum_{r=1}^{k} \frac{1}{n_r} \sum_{i,j \in C_r} \text{sim}(i,j) \quad \cdots\cdots\cdots\cdots (1)$$

여기서, n_r은 Cr 내의 데이터 개수, k는 클러스터 개수

계층적 클러스터링 알고리듬의 단계는 그림 4와 같다.

```
Step 1. Initialize
    For all i
      Ci ← Si ∈ D

Step 2. Compute the value of
          the criterion function
    For each Ci, Cj ∈ D
      compute the criterion function (Eq.(1))

Step 3. Merge
    Cnew ← merge (Ci, Cj) for the max
      value of the criterion function (Eq.(1))
Step 4. Check the condition
      If ( |Ci| > k ) then go to Step 2.
       Else go to Step 5.

Step 5. Exit
```

그림 4. 계층적 클러스터링 알고리듬
Fig. 4. Hierarchical clustering algorithm

Step 1은 초기화 단계로서 데이터베이스 D를 액세스하여 각각의 데이터를 하나의 클러스터로 설정한다. Step 2는 두 클러스터가 합병이 될 경우의 평가함수 식(1)의 값을 구하는 단계로, 현재 n개의 클러스터가 있다고 하면, n·(n-1)/2개의 평가함수 값을 계산한다. Step 3은 합병 단계로서,

Step 2에서 계산한 평가함수 값들 중 가장 큰 값을 주는 두 개의 클러스터를 합병한다. Step 4는 조건 검사 단계로서 클러스터의 개수가 지정된 클러스터 개수보다 크면 Step 2로 간다. 그렇지 않으면 Step 5로 간다. 마지막으로, Step 5는 종료 단계로서 알고리듬을 끝낸다.

데이터 전처리 과정을 거친 데이터들은 결정 변수 값들이 없는 데이터들로 이루어져 있다. 본 논문에서는 이들 데이터들을 계층적 클러스터링 알고리듬을 적용하여 k 개의 그룹으로 클러스터링 한 후, 이 결과를 결정 변수의 값으로 사용한다.

4. 러프셋 이론을 이용한 리덕트 생성

전 단계에서 도출된 데이터들로부터 직접 규칙을 추출하기에는 데이터들의 집합이 너무 크므로, 이번 단계에서는 데이터들의 크기를 줄이기 위한 과정을 처리한다. 이를 위해 러프셋 이론의 리덕트를 이용한다. 리덕트를 이용하여 데이터셋의 크기를 줄이기 위한 연구로는 Questier et. al.[14]과 Ohrn[15]이 있다.

의사결정 테이블은 하나 이상의 리덕트들을 가질 수 있는데, 이중에 하나의 리덕트가 테이블을 대신하여 사용될 수 있다. 의사결정 테이블에서 모든 리덕트들을 찾아내는 것은 NP-Hard 문제이지만, 많은 응용 분야에 있어 이들 모두를 찾아낼 필요 없이 하나의 리덕트면 충분하다.

본 논문에서는 Ohrn[15]가 제안한 방법을 통하여 리덕트들을 찾는다. 여기서는 히팅 셋이라는 용어를 정의하고, 최소 히팅 셋을 계산하기 위하여 유전자 알고리듬을 이용하며, 여기서 사용되는 적합도 함수 f는 다음과 같이 정의된다.

$$f(B) = (1-\alpha) \times \frac{cost(A) - cost(B)}{cost(A)} + \alpha \times \min\left\{\epsilon, \frac{|\,[S \in S \mid S \cap B \neq \varnothing]\,|}{|\,S\,|}\right\} \quad \text{(2)}$$

5. 리덕트를 이용한 규칙 생성

본 단계에서는 전 단계에서 찾아낸 리덕트를 이용하여 규칙을 생성하는 단계이다. 규칙을 생성하기 위해서는 다음과 같은 네 가지 휴리스틱 방법을 이용한다.

1) 같은 결정 변수를 갖는 규칙들끼리 클러스터링한다.
2) 같은 결정 변수를 갖는 속성 값들을 속성 값들의 범위로 변환한다.
3) 규칙을 생성하기 위해 사용되지 않는 속성들은 제거한다.
4) 속성들과 결정 변수들을 이용하여 의사결정 테이블을 if-then 형태의 규칙으로 표현한다.

그림 5와 같이 세 가지 속성들로 이루어진 규칙들이 있다고 하자.

	A1	A2	A3	D
1	b	f	1	high
2	a	f	2	medium
3	c	f	3	low
4	b	f	2	high
5	c	f	4	low

그림 5. 규칙들의 예
Fig. 5. Example of rules

그림 5로부터 규칙 생성 1)단계를 적용하면 다음과 같은 그림 6을 얻는다.

	A1	A2	A3	D
1	b	f	1	high
4	b	f	2	high
3	c	f	3	low
5	c	f	4	low
2	a	f	2	medium

그림 6. 그림 5로부터 나온 규칙들
Fig. 6. Patterns of rules from Fig. 5

그림 6으로부터 규칙 생성 2)단계를 적용하면 다음과 같은 그림 7을 얻는다.

	A1	A2	A3	D
1,4	b	f	1-2	high
3,5	c	f	3-4	low
2	a	f	2	medium

그림 7. 그림 6으로부터 나온 규칙들
Fig. 7. Patterns of rules from Fig. 6

그림 7로부터 규칙 생성 3)단계를 적용하면 다음과 같은 그림 8을 얻는다.

	A1	A3	D
1,4	b	1-2	high
3,5	c	3-4	low
2	a	2	medium

그림 8. 그림 7로부터 나온 규칙들
Fig. 8. Patterns of rules from Fig. 7

그림 8로부터 나온 규칙들을 if-then 형태의 규칙들로 변환하면 그림 9와 같다.

```
if A1='b' and A3=(1-2)
then D='high'

if A1='c' and A3=(3-4)
then D='low'

if A1='a' and A3=(2)
then D='medium'
```

그림 9. 그림 8로부터 나온 규칙들
Fig. 9. Patterns of rules from Fig. 8

Ⅳ. 실험 결과

이번 장에서는 본 논문에서 제안하는 방법에 대한 실험결과를 제시한다. 먼저 본 실험에 사용된 데이터 셋은 UCI KDD 아카이브[16]에 포함되어 있는 데이터 셋으로서, Soybean과 Zoo 데이터 셋이다.

표 1에서 보는 바와 같이 Soybean 은 47개의 레코드와 35개의 속성을 가지며, Zoo 는 101개의 레코드와 17개의 속성으로 구성되어 있다. Soybean 데이터 셋은 잎의 상태, 줄기의 상태, 씨앗의 상태 등 35개의 속성들을 가지고 콩에 생기는 병의 종류를 분류한 데이터 셋이다. Zoo 데이터 셋은 다리의 개수, 꼬리의 유무, 머리카락의 유무, 척추의 유무 등 17개의 속성들로 동물 타입을 분류한 데이터 셋이다.

표 1. 데이터 셋
Table 1. Data set

데이터 셋	데이터 개수	속성 개수
Soybean	47	35
Zoo	101	17

실험은 표 1의 데이터 셋을 본 논문에서 제안하는 방법을 사용하여 수행하였으며, 표 2와 같은 결과를 얻었다. 표 2에서 k는 클러스터의 개수이다.

표 2. 실험 결과
Table 2. Experimental Results

데이터 셋		리덕트 개수	규칙 개수
Soybean	k=2	2	5
	k=3	2	5
	k=4	3	30
	k=5	3	34
Zoo	k=2	2	4
	k=3	2	10
	k=4	2	10
	k=5	3	15

실험에서 얻어진 규칙의 샘플들은 다음의 그림 10과 같다.

```
if milk=1 and legs=4 and tail=0
then decicion = 2

if milk=1 and legs=4 and tail=1
then decicion = 2

if milk=0 and legs=0 and tail=1
then decicion = 1

if milk=0 and legs=2 and tail=1
then decicion = 4
```

그림 10. 생성된 규칙들의 예
Fig. 10. Example of extracted rules

그림 10의 규칙들은 Zoo 데이터 셋의 경우이며, k=5로 클러스터링 한 후의 리덕트가 {milk, legs, tail} 속성들로 구성된 경우이다. 예를 들면, 그림 10의 첫 번째 규칙은 '젖을 먹으며, 다리가 4개고, 꼬리가 없으면, 이런 류의 동물은 타입이 2번째인 동물이다'라는 규칙을 표현한다.

Ⅴ. 결론

데이터 마이닝이란 전산 지능 분야의 새로운 영역중 하나이며, 기계학습, 클러스터 분석, 회귀 분석, 뉴럴 네트워크 등과 관련이 있는 분야이다. 최근에는 대량의 데이터들이 디지털 형태로 제공됨에 따라 이 분야에 대한 연구가 활발해 지고 있다.

본 논문에서는 클러스터링 알고리듬과 러프 셋 이론을 이용하여 데이터들로부터 규칙을 추출하는 통합적인 방법을 제안하였다. 이러한 방법은 기존의 데이터 마이닝 기법들을 개별적으로 적용하는 것이 아니라, 각 개별 기법들을 통합화하여 자동적으로 규칙을 생성하는 방법이다. 이런 데이터 규칙 추출 방법론은 인간이 대량의 데이터들을 분석하여 이들 정보로부터 성공적인 의사결정을 내리는데 도움을 줄 수 있기에 많은 분야에 이용될 수 있다.

본 연구에서는 UCI KDD 아카이브에 포함되어 있는 Soybean과 Zoo 데이터 셋만을 대상으로 실험을 수행하였는데, 향후에는 본 논문에서 제안하는 방법을 UCI KDD 데이터 셋 외에 다양한 데이터 셋에 적용해 보는 것이 필요하겠다. 특히 웹 로그 데이터 등을 이용하여 고객관계관리(CRM)나 부정사용방지 시스템(fraud detection)에 활용해 보는 것이 필요하겠다.

참고문헌

〔1〕 J. Han, M. Kamber, Data Mining: concepts and techniques, Morgan Kaufmann publishers, 2000.

〔2〕 오승준, "확장된 시퀀스 요소 기반의 유사도를 이용한 계층적 클러스터링 알고리즘", 한국컴퓨터정보학회 논문지, 제11권, 제5호, 2006.

〔3〕 오승준, "범주형 시퀀스 데이터의 K-Nearest Neighbour 알고리즘", 한국컴퓨터정보학회 논문지, 제10권, 제2호, 2005.

〔4〕 Z. Pawlak, "Rough sets", Int. J. Comput. Inform. Sci. Vol. 11, pp 341-356, 1982.

〔5〕 Z. Pawlak, Rough sets: theoretical aspects of reasoning about data, Kluwer Academy Publisher, 1991.

〔6〕 D. Kim, "Data classification based on tolerant rough set", Pattern Recognition, Vol. 34 No.8, pp.1613-24, 2001.

〔7〕 T. McKee, T., Lensberg, "Genetic programming and rough sets: a hybrid approach to bankruptcy classification", European Journal of Operational Research, Vol. 136, No.2, pp.436-51, 2002.

〔8〕 A. Kusiak, "Rough set theory: a data mining tool for semiconductor manufacturing", IEEE Transactions on Electronics Packaging Manufacturing, Vol. 24, No.1, pp.44-50, 2001.

〔9〕 K. Thangavel, Q. Shen, A. Pethalakshmi, "Application of clustering for feature selection based on rough set theory approach", AIML Journal, Vol. 6, No. 1, 2006.

〔10〕 S. Asharaf, M. N. Murty, S. K. Shevade, "Rough set based incremental clustering of interval data", Pattern Recognition Letters, Vol. 27, pp.515-519, 2006.

〔11〕 A. Kusiak, J. A. Kern, K. H. kernstine, and B. T. L. Tseng, "Autonomous Decision-Making: A Data Mining Approach", IEEE Transaction on Information Technology in Biomedicine, Vol. 4, No. 4, 2000

〔12〕 H. Sakai, K. Kobe, and M. Nakata, "Rough Sets Based Rule Generation from Data with Categorical and Numerical Values", Journal of Advanced Computational Intelligence and Intelligent Informatics, Vol. 12, No. 5, 2008

〔13〕 J. Dougherty, R. Kohavi, and M. Sahami, "Supervised and unsupervised discretization of continuous features", Proc. 13th Int. Conf. on Machine Learning, pp 194-202, 1995.

〔14〕 F. Questier, I. Arnaut-Rollier, B. Walczak, and D.L. Massart, "Application of rough set theory to feature selection for unsupervised clustering", Chemometrics and Intelligent Laboratory Systems, Vol. 63, 2002.

〔15〕 A. Ohrn. "Discernibility and rough sets in medicine: tools and applications", PhD thesis, Norwegian Univ. of Science and technology, 1999.

〔16〕 C. L. Blake and C. J. Merz. UCI Repository of Machine Learning Databases, 1998.

저 자 소 개

오 승 준

2004년 8월 한양대학교
산업공학과. 공학박사
2005~ 현재 :
경기공업대학
산업경영과 교수

박 찬 웅

1997년 2월 한양대학교
산업공학과. 공학박사
1997~ 현재 :
경원대학교
산업정보시스템공학과 교수