韓國컴퓨터情報學會 論文誌
第17卷 第3號, 2012. 3.

2012-17-3-4-3

입자개념계층구조를 기반으로 하는 데이터 분석 기법

강 유 경*, 황 석 형, 김 응 희, 엄 태 정

A Study on Data Analysis Approach based on Granular Concept Hierarchies

Yu-Kyung Kang*, Suk-Hyung Hwang, Eung-Hee Kim, Tae-Jung Eom

요 약

본 논문에서는 형식개념분석기법에 입자의 정밀도를 조절하기 위해 스케일링 정도(Scaling level)를 도입하여 다양한 관점과 추상화 수준을 토대로 데이터를 분류하는 새로운 기법을 제안하였다. 이 기법의 특징은 주어진 데이터를 다양한 기준에 맞춰서 입자화하고, 입자들 사이에 관계를 토대로 분석하여, 입자개념계층구조(Granular Concept Hierarchy)를 구축함으로써, 데이터를 분석하고자 하는 사용자의 의도 또는 목적에 맞추어서 다양한 분류가 가능하다는 것이다. 또한, 본 연구에서 제안한 기법을 지원하는 도구(G-Tool)를 개발하였으며, 본 연구에서 제안한 기법의 유용성을 검토하기위해 실제 데이터를 대상으로 G-Tool을 사용하여 실험을 실시하였으며, 그 결과 사용자의 목적에 맞는 다양한 형태로 데이터를 분류할 수 있음을 확인하였다. 기존의 형식개념분석기법에는 입자의 정밀도를 조절할 수 없어서 특정한 어느 한 관점에 대한 분류만 가능하였으나, 본 연구에서 제안한 기법은 사용자의 의도 또는 목적에 맞추어서 다양한 종류의 스케일 정보를 조합하고 스케일링 정도를 조절함으로써 다양한 관점을 반영한 다양한 분류가 가능하다.

▸ Keyword : 입자화 컴퓨팅, 형식개념분석, 개념적 분류, 입자개념계층구조

Abstract

In this paper, we propose a novel data analysis approach that extracts granules suitable for various perspectives by introducing scaling level into formal concept analysis in order to control the level of granularity. Based on our approach, we can extract various granules from the given data set and constructs granular concept hierarchies based on the relations between the granules. Therefore, we can classify the given data with respect to the purpose or the intention of user's viewpoints. And, we developed G-Tool that supports our approach. In order to verify the usefulness

• 제1저자 : 강유경 • 교신저자 : 강유경
• 투고일 : 2011. 11. 09, 심사일 : 2011. 11. 21, 게재확정일 : 2011. 12. 23.
* 선문대학교 컴퓨터공학과(Dept. of Computer Engineering, Sunmoon University)

of our proposed approach and G-Tool, we have done some experiments for real data set and reported about results of our experiments. From the experiments' results, we can verify our approach with G-Tool can be useful and suitable for classifying the given data with various scaling levels. The traditional formal concept analysis cannot control the level of granularity and can only classify for a particular perspective. However, our proposed approach can classify the given data with respect to user's purpose or intention by combining of diverse scale information and scaling levels.

▸ Keyword : Granular Computing, Formal Concept Analysis, Conceptual Classification, Granular Concept Hierarchy

I. 서 론

입자화 컴퓨팅 기법(Granular Computing Approach)[1-3]은 주어진 데이터로부터 입자(granule)를 기본 단위로 하여 유용한 정보를 추출하는 데이터분석기법의 일종으로서, 최근 다양한 분야에서 널리 사용되고 있다. 하나의 입자는 주어진 도메인에 포함된 객체들의 부분집합 중의 하나를 의미하며, 주어진 도메인 안에 모든 객체들을 입자들로 분류하는 것을 입자화(granulation)라 부른다. 주어진 도메인에 대한 입자화는 객체들을 클러스터들로 그룹핑(grouping)하거나 도메인의 객체들을 부분집합들로 분할하는 것으로, 데이터를 분석하고자하는 관점에 따라 다양한 입자들이 생성될 수 있으며, 일반적으로 이러한 입자화 컴퓨팅 공정은 주어진 대량의 데이터에 대한 그룹핑, 파티션(partition), 군집화(clustering), 분류(classification) 등의 기법을 적용하여 분류체계를 구축하는데 필수적이다. 입자화 컴퓨팅기법을 기반으로 하는 데이터 분석 분야에는 형식개념분석기법(FCA : Formal Concept Analysis), 퍼지집합분석기법(FSA : Fuzzy Set Analysis), 러프집합분석기법(RSA : Rough Set Analysis) 등이 있다[1-7].

형식개념분석기법[4]은 주어진 데이터로부터 공통속성을 갖는 객체들을 클러스터링하여 정보의 최소단위로써 개념(Concept)들을 추출하고 그들 사이의 관계를 토대로 계층화하여 데이터에 내재된 개념들의 구조를 가시화 해주는 입자화 컴퓨팅의 한 종류이다. 퍼지집합분석기법[5]은 '네' 또는 '아니오' 등 이분법으로는 나타내기 힘든 실세계의 데이터를 정량적으로 표현하는 퍼지집합으로부터 각 원소가 그 집합에 귀속되어지는 정도를 0부터 1사이의 수로 나타낸 귀속도(歸屬度, membership degree)를 토대로 공통속성을 갖는 객체들을 분류함과 동시에 각 속성을 어느 정도 포함하고 있는지에 관한 정량적인 수치 또한 분석할 수 있다. 러프집합분석기법[5]은 실세계의 애매모호한 데이터를 다루기 위해서, 어떤 집합에 확실하게 분류되는 하한근사(Lower Approximation)와 불확실하게 분류되는 상한근사(Upper Approximation)를 사용하여 구별하기 애매모호한 원소들을 하나의 개념으로 포함시켜서 추출함으로써, 주어진 불완전한 데이터를 분류하는 기법이다.

위와 같은 분석기법들은 주로 입력된 데이터를 특정 관점에 맞춰서 분석하여 입자들을 추출하는 방식으로 데이터를 분석하고 있으나, 다양한 목적에 맞도록 적절히 입자의 정밀도를 조절하면서 데이터를 분석하지 못한다. 따라서, 본 논문에서는 형식개념분석기법에 입자의 정밀도를 조절하기 위해 스케일링 정도(Scaling level)를 도입하여 다양한 관점과 추상화 수준을 토대로 하여, 주어진 데이터를 다양한 기준에 맞춰서 입자화하고, 입자들 사이에 관계를 토대로 분석하여, 입자개념계층구조(Granular Concept Hierarchy)를 구축함으로써, 데이터를 분류하는 기법을 제안하였다. 또한, 본 연구에서 제안한 기법을 지원하는 도구(G-Tool)를 개발하였으며, 본 연구에서 제안한 기법의 유용성을 검토하기위해 실제 데이터를 대상으로 G-Tool을 사용하여 실험을 실시하고 그 결과를 기술하였다.

본 논문은 다음과 같이 구성된다. 제2장에서는 대량의 데이터에 대한 분류체계구축과 관련된 기존의 연구들에 대하여 설명한다. 제3장에서는 다양한 기준으로 데이터를 분류하기 위한 스케일 조합 트리(SCT : scale combination tree)를 토대로, 입자개념계층구조의 구축에 대하여 설명한다. 제4장에서는, 본 연구에서 제안한 기법을 지원하는 도구 G-Tool의 개발에 대하여 소개하고, 실제 데이터를 대상으로 도구를 사용하여 실험한 결과를 설명한다. 제5장에서는 결론과 향후 연구과제에 대해서 설명한다.

II. 관련 연구

본 장에서는, 주어진 데이터를 다양한 방법으로 분류하고,

분류된 데이터들 사이에 관계를 파악하여 분류체계를 구축하기 위한 관련 연구들([1, 4, 8, 9])에 대하여 요약하였다.

형식개념분석기법에서는[4] 주어진 도메인으로부터 객체(Object)와 속성(Attribute)들을 추출하고, 이들 사이의 포함관계를 이진데이터 테이블 형태인 Formal Context로 표현한다. Formal Context K = (G, M, I)는 객체들(Objects)의 집합 G와 속성들(Attributes)의 집합 M, 그리고 G와 M사이의 이항관계 I⊆G×M로 구성된다. 주어진 Formal Context로부터 Galois Connection(ext:$2^M \rightarrow 2^G$와 int:$2^G \rightarrow 2^M$)을 기반으로 공통속성을 갖는 객체들을 (O, A)와 같은 형태의 개념들로 추출할 수 있다(단, O⊆G, A⊆M일때, (O, A)⇔(int(O)=A∧ext(A)=O)). 주어진 Formal Context K=(G, M, I)로부터 추출된 모든 개념들의 집합 C와 그들 사이의 상·하위개념관계≤를 기반으로 개념계층구조 L:=(C, ≤)를 구축할 수 있다. 개념계층구조를 구축함으로써, 주어진 데이터를 객체단위로 분류하고 체계화할 수 있으며, 개념계층구조를 토대로 데이터 분류 및 연관규칙 등 유용한 정보를 추출할 수 있다.

형식개념분석기법에서는 복잡한 구조를 갖는 대량의 데이터를 다루기 위해서 개념적인 스케일(conceptual scales)이 개발되었고, 개념적인 스케일은 데이터를 개념적으로 구조화하는 메타데이터로 여겨졌다. 그러나, 방대한 양의 데이터와 관련 지식을 필요로 하는 응용분야에서는 개념적인 스케일의 수 또한 크게 증가하여, 다양한 수준의 개념 표현 및 설정을 사용자에게 지원하는 방법들이 필요하게 되었다. G. Stumme[8]는 이러한 문제를, 기존의 평면적인 형태의 스케일들을 구조적으로 계층화된 스케일로 확장하고 중첩된 스케일링이라 불리는 가시화 기법을 제안함으로써 해결하고자 하였다. G. Stumme는 주어진 도메인의 분류체계(taxonomy)로부터 고수준의 일반화된 스케일들(higher level scales)을 추출하여, 이를 토대로 데이터를 분석함으로써, 주어진 데이터에 대한 조금 더 일반적인 수준의 정보를 제공하고자 하였다.

입자화 컴퓨팅[1]을 기반으로 하는 데이터 분석 분야에서 파티션은 대상이 되는 객체집합 U를 입자화하는데 사용되는 가장 일반적인 방법이다. 즉, 파티션은 주어진 데이터를 분류하는 대표적인 데이터 마이닝의 기법이다. 객체들의 집합 U에 대한 파티션 π = {X_i | 1≤ i ≤n}은 공통 원소를 갖지 않는 U의 부분집합 n개로 구성된 집합으로, 공집합을 원소로 갖지 않으며, 파티션 π의 모든 원소를 합집합하면 전체집합 U가 나오게 되는 특징을 갖고 있다. 다양한 기준에 따라 대상이 되는 객체들의 집합 U에 대한 다양한 파티션들이 생성될 수 있다.

형식개념분석을 적용하여 추출된 개념들의 수가 많으면 개념계층구조가 너무 상세화되어 복잡해지고, 추출된 개념들의 수가 적으면 개념계층구조가 너무 추상화되어 데이터를 해석하는데 어려움이 있다. 이와 같은 문제점을 해결하고자 R. Belohlavek[9]는 입자화 컴퓨팅 분야에서 사용되는 입자화 개념을 도입하여, 주어진 데이터의 각 속성에 대한 입자화 정도 트리(granularity level-tree)를 구축하고, 이를 토대로 추출되는 개념들의 입자화 정도를 조절하는 방법을 제안하였다. 즉, 구축된 입자화 정도 트리의 레벨에 따라 주어진 데이터에 대하여 다양한 분류체계를 구축할 수 있다.

III. 입자개념계층구조의 구축

본 장에서는 분석 대상 객체들의 스케일링 수준을 나타내는 스케일 통합 트리(SCT) 구축을 위한 제반 정의들과 SCT를 기반으로 하는 입자개념계층구조의 구축에 대하여 기술한다.

3.1 제반정의

형식개념분석기법과 같은 데이터분석기법에서 분석대상이 되는 데이터는 여러 가지 다양한 값을 가지는 속성들과 객체들, 그리고 객체와 속성 사이의 관계를 나타내는 테이블형태로 정리되어 표현할 수 있다[1,4].

[정의1] 데이터 테이블 S = (G, M, W, I)는 다음과 같은 요소들로 구성된다[4]:

- 객체들의 집합 G = {g_1, g_2, …, g_m},
- 속성들의 집합 M = {m_1, m_2, …, m_n},
- 속성이 갖는 값들의 집합 W = {W_m | m ∈ M},
- I(g, m) : G×M → W_m은 다음의 조건을 만족하는 G×M로부터 W_m로의 사상. 즉,

$$I(g, m)\in W_m,\ \forall g\in G,\ m\in M$$ ■

즉, G와 M의 원소는 각각 해당 테이블의 객체들과 각 객체들이 가질 수 있는 속성들, 그리고 그 속성의 값들을 나타낸다. 또한, I(g, m) : G×M → W_m은 어떤 객체 g가 속성 m을 가지고 있고 그 속성의 값이 w∈W_m임을 나타낸다. 임의의 속성 m∈M과 모든 객체 g∈G에 대하여 사상I(g, m)에 의해 대응되는 속성값 w가 존재 한다면, “속성 m은 완전(complete)하다”라고 부른다. 데이터 테이블 S의 모든 속성 m∈M이 완전한 경우, “데이터 테이블 S는 완전한 테이블(Complete Data Table)이다”라고 부른다. 이 논문에서는 데이터 테이블 S는 완전한 것만을 대상으로 한다.

표 1. 데이터 테이블
Table 1. Data Table

객체 \ 속성	키	몸무게	나이
O1	188	103	22
O2	197	120	19
O3	143	38	15
O4	136	46	32
O5	182	97	53

표1은 데이터 테이블의 예로써, 5명의 사람에 대한 "키", "몸무게", "나이"에 대한 정보를 나타낸다. 이와 같은 다양한 값을 갖는 데이터 테이블로부터 개념들을 추출하고 개념계층구조를 구축하기 위해서는 특정한 규칙에 따라 주어진 데이터를 이진데이터로 변환할 필요가 있다. 이러한 변환과정을 스케일링(Scaling)이라고 한다. 스케일링하기 위해서, 주어진 데이터의 각 속성들은 스케일 테이블(해석기준)을 토대로 이진데이터로 변환되며, 다양한 값을 갖는 데이터 테이블의 각 속성들은 스케일 테이블에 의해서 해석된다.

[정의2] 다양한 값을 갖는 데이터 테이블 S의 속성 m∈M에 대한 스케일 테이블 S_m = (W_m, M_m, I_m)은 데이터 테이블 S에서 속성 m이 갖는 값들의 집합 W_m과 속성m을 세분화하여 표현한 속성들의 집합 M_m, 그리고 W_m과 M_m사이의 이항관계 $I_m \subseteq W_m \times M_m$로 구성된다[4]. ■

표1에서 속성 "키"에 대한 스케일 테이블 $S^1_{키}$는 표2와 같다.

표 2. 속성 "키"에 대한 스케일 테이블 $S^1_{키}$

Table 2. Scale Table $S^1_{키}$ for attribute "height"

$S^1_{키}$	매우 크다	크다	작다	매우 작다
188		X		
197	X			
143			X	
136				X
182		X		

스케일 테이블은 주어진 데이터 테이블을 분석하고자 하는 목적 또는 관점에 따라서 다양하게 여러 개의 스케일 테이블들이 만들어질 수 있다. 주어진 데이터 테이블 S= (G, M, W, I)의 임의의 속성 m∈M에 대한 다양한 관점에서 정의되는 스케일 테이블들의 집합 SS(m)은, 다음과 같이 정의된다.

$$SS(m) = \{S^1_m, S^2_m, ..., S^k_m\},$$

단, $S^i_m = (W_m, M^i_m, I^i_m)$는 속성m에 대한 스케일 테이블이다.

표 3. 속성"키"에 대한 스케일 테이블 $S^2_{키}$

Table 3. Scale Table $S^2_{키}$ for attribute "height"

$S^2_{키}$	장신	단신
188	x	
197	x	
143		x
136		x
182	x	

표3은 주어진 데이터 테이블(표1)의 속성 "키"에 대해 표2와는 다른 기준으로 데이터를 해석하기 위한 스케일 테이블 $S^2_{키}$이다. 표1의 속성 "키"에 대한 스케일 테이블들의 집합 SS(키) = {$S^1_{키}$, $S^2_{키}$}이다.

[정의3] 주어진 데이터 테이블 S = (G, M, W, I)의 임의의 속성 m에 대한 스케일 테이블 S^i_m와 S^j_m에 대하여(단, $S^i_m, S^j_m \in$SS(m)), 상•하위관계($\leq$: SS(m)×SS(m)) $S^i_m \leq S^j_m$는 다음과 같이 정의된다.

$$S^i_m \leq S^j_m \Leftrightarrow ① W^i_m = W^j_m$$

$$② \forall a \in M^i_m, \exists b \in M^j_m [ext(a) \subseteq ext(b)]$$

$$③ \exists \beta : M^i_m \rightarrow M^j_m [\forall w \in W^i_m, a \in M^i_m$$

$$[(w,a) \in I^i_m \Rightarrow (w, \beta(a)) \in I^j_m]]$$

단, $ext(a) = \{w \in W_m \mid \forall a \in \dot{M}^i_m : (w,a) \in I^i_m\}$ ■

주어진 데이터 테이블의 임의의 속성m에 대한 스케일 테이블 S^i_m와 S^j_m는 상•하위관계 $S^i_m \leq S^j_m$이며, 다음과 같은 조건을 만족한다.

① 하위 스케일 테이블 S^i_m의 객체집합과 상위 스케일 테이블 S^j_m의 객체집합은 같다. 즉, S^i_m와 S^j_m는 동일한 객체집합을 대상으로 서로 다른 해석기준을 표현해 놓은 것이다.

② 하위 스케일 테이블 S^i_m에서 임의의 속성을 갖는 객체들의 집합은 상위 스케일 테이블 S^j_m에서 임의의 속성을 갖는 객체들의 집합에 포함된다. 즉, 하위 스케일 테이블의 속성들과 상위 스케일 테이블의 속성들은 서로 다르지만, 각각의 속성들에게 대응하는 객체집합은 상위 스케일 테이블의 객체들집합이 하위 스케일 테이블의 객체들집합을 포함한다.

③ 하위 스케일 테이블 S_m^i 의 이항관계는 맵핑β에 의해 상위 스케일 테이블 S_m^j 의 이항관계로 맵핑된다. 즉, 하위 스케일 테이블에 존재하는 이항관계는 정보의 손실 없이 상위 스케일 테이블에 반드시 존재한다.

예를 들어, 주어진 데이터 테이블(표2)의 속성 "키"에 대한 스케일 테이블들 $S_{키}^1 = (W_{키}^1, M_{키}^1, I_{키}^1)$ 과 $S_{키}^2 = (W_{키}^2, M_{키}^2, I_{키}^2)$ 사이에는 상•하위관계(≤) $S_{키}^1 \leq S_{키}^2$ 가 존재한다. 즉,

① $W_{키}^1 = W_{키}^2 = \{188, 197, 143, 136, 182\}$

② $M_{키}^2$ ={장신, 단신},

$M_{키}^1$ ={매우 크다, 크다, 작다, 매우 작다}

- ext({매우 크다})={197} ⊆ {188,197,182}=ext({장신})
- ext({크다})={188, 182}⊆ {188, 197, 182}=ext({장신})
- ext({작다}) = {143} ⊆ {143, 136} = ext({단신})
- ext({매우 작다})= {136} ⊆ {143, 136} = ext({단신})

③ $I_{키}^2$ ={(188, 장신), (197, 장신), (182, 장신), (143, 단신), (136, 단신)}

$I_{키}^1$ ={(188,크다), (197,매우 크다), (182,크다), (143,작다), (136, 매우 작다)}

- (188, 크다) ⇒ (188, 장신)
- (197, 매우크다) ⇒ (197, 장신)
- (182, 크다) ⇒ (182, 장신)
- (143, 작다) ⇒ (143, 단신)
- (136, 매우작다) ⇒ (136, 단신)

이 논문에서 다루는 스케일 테이블들은 정의2에서와 같이 동일한 객체들에 대하여 임의의 속성m에 대한 다양한 해석 기준을 제공하고 있다. 따라서, 상•하위관계를 갖는 각 스케일 테이블의 객체집합(W_m^i 와 W_m^j)는 동일하다.

이 논문에서는 일반성을 상실하지 아니하고, 이후 논의 대상이 되는 SS(m)의 각 구성요소들은 모두 상•하위관계를 갖는 것으로 가정한다. 즉, $SS(m) = \{S_m^1, S_m^2, ..., S_m^k\}$ 일 때, 모든 i, j(단, i < j)에 대해서, $S_m^i \leq S_m^j$ 으로 가정한다.

주어진 데이터 테이블의 속성 m에 대한 스케일 테이블들의 집합 SS(m)의 각 요소들(스케일 테이블)과 정의3의 상•하위관계를 토대로 스케일 조합 트리(Scale Composition Tree)를 구축할 수 있다.

[정의4] 데이터 테이블 S = (G, M, W, I)와 임의의 속성 m∈M에 대한 스케일 테이블들의 집합 $SS(m) = \{S_m^1, S_m^2, ..., S_m^k\}$ (단, $S_m^i = (W_m, M_m^i, I_m^i)$)이고, 임의의 S_m^i 와 S_m^j (i ≤ j)에 대하여 $S_m^i \leq S_m^j$ 가 성립한다.)이 주어졌을 때, 속성 m에 대한 스케일 조합 트리 SCT(m)은, 다음과 같이 정점m을 뿌리로 갖는 트리(rooted tree)이다.

SCT(m) = (m:V, E),

- $V = \{m\} \bigcup \dot{M}_m^i \bigcup W_m^i$:정점들의 집합단, $\dot{M}_m^i = \{i\} \times M_m^i$ 이다)
- E⊆V×V : 변들의 집합

$$(1) E = E \bigcup \{(m, x) \mid \forall x \in \dot{M}_m^k\}$$

$$(2) E = E \bigcup \{(x, y) \mid \forall y \in \dot{M}_m^{i-1} \exists x \in \dot{M}_m^i : ext(y) \subseteq ext(x), \text{ 단}, 2 \leq i \leq k\}$$

$$(3) E = E \bigcup \{(y, z) \mid (z, y) \in I_m^1\}$$ ■

SCT(m)는 정점들의 집합 V와 변들의 집합 E로 구성된다. 정점들의 집합V는 루트인 m과 SS(m)의 각 요소들(스케일 테이블)의 속성들, 그리고 객체들로 구성되며, 변들의 집합 E는 정점과 정점의 쌍으로 구성된다. (1)은 루트 m과 최상위 스케일 테이블의 속성을 연결하는 변을 나타내고, (2)에서는 상•하위관계에 있는 SS(m)의 요소들(스케일 테이블)의 속성들을 추출하여 하위 스케일 테이블의 속성을 갖는 객체들과 상위 스케일 테이블의 속성을 갖는 객체들 사이의 포함관계에 의해 변을 E에 추가한다. (3)에서는 최하위 스케일 테이블에 정의되어 있는 객체와 속성 간의 이항관계 I_m^1 의 요소 (z, y)를 현재 구성하고 있는 SCT(m)의 변의 집합 E에 추가한다.

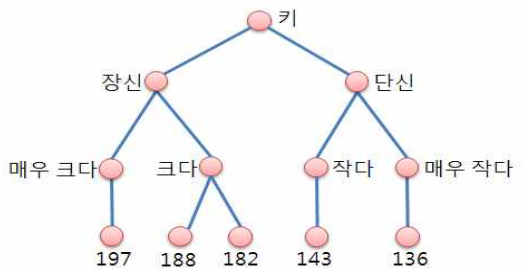

그림 1. 표1의 속성 "키"에 대한 스케일 조합트리 SCT(키)
Fig. 1. Scale Combination Tree SCT(height) for attribute "height" of the table 1.

표1의 속성 "키"에 대한 스케일 테이블들의 집합 SS(키) = {$S^1_{키}$, $S^2_{키}$}를 토대로 구축된 스케일 조합 트리 SCT(키)는 다음과 같이 구성된다(그림1 참조).

SCT(키) = (키:V, E)

V = {키, 장신, 단신, 매우 크다, 크다, 작다, 매우 작다, 197, 188, 182, 143, 136},

E = {(키, 장신), (키, 단신), (장신, 매우 크다), (장신, 크다), (단신, 작다), (단신, 매우 작다), (매우 크다, 197), (크다, 188), (크다, 182), (작다, 143), (매우 작다, 136)}.

이때, SCT(m)의 임의의 정점 v_1, v_n에 대하여, v_1에서 v_n까지의 경로 P(v_1, v_n)은 <v_1, v_2, ... v_n>으로 구성되는 정점들의 열(sequence)이다. 단, $v_i \in V(1 \le i \le n)$이고, $v_i \neq v_j (i \neq j)$이며, (v_i, v_{i+1})$\in$ E(i=1, 2, ..., n-1)이다.

속성 m에 대한 스케일 조합 트리 SCT(m) = (m:V, E)가 주어졌을 때, 경로의 길이 |P(x, y)|는 경로 P(x, y)상의 변의 개수를 나타낸다(단, |P(x, x)| = 0). 스케일 조합 트리 SCT(m)의 임의의 정점 v∈V의 스케일링 수준(scaling level) sLevel(m, v)는 m으로부터 v까지의 경로의 길이를 나타내다. 즉, sLevel(m, v) = |P(m, v)|이다. 또한, 속성 m에 대한 스케일 조합 트리 SCT(m) = (m:V, E)가 주어졌을 때, sLevel k에 존재하는 정점들의 집합 SLV(m, k)은 다음과 같이 정의 된다:

SLV(m, k) = {x∈V | sLevel(m, x) = k}.

그림1과 같이 속성 "키"에 대한 스케일 조합 트리에서 sLevel 0, 1, 2, 3 각각에 대한 SLV는 다음과 같다.

(1) SLV(키, 0) = {키}

(2) SLV(키, 1) = {장신, 단신}

(3) SLV(키, 2) = {매우 크다, 크다, 작다, 매우 작다}

(4) SLV(키, 3)= {197, 188, 182, 143, 136}.

[정의5] 주어진 SCT(m)에 대하여, 임의의 sLevel k의 정점집합 SLV(m, k)을 기준으로 하는 파티션 π(m, k)은 다음과 같이 정의한다.

$$\pi(m,k) = \bigcup_{x \in SLV(m,k)} OS(x),$$

(단, $OS(x) = \{g \in G \mid \forall x \in \dot{M}^i_m : I(g,m) = w \wedge (w,x) \in I^i_m\}$)

또한, PS(m)은 속성 m에 대한 스케일 조합트리의 모든 레벨에 대한 모든 파티션들의 집합으로 다음과 같이 정의한다.

$$PS(m) = \bigcup_{i=0..k} \pi(m,i)$$ ■

이때, 파티션 π(m, k)는 데이터 테이블 S의 객체집합 U에 대한 파티션{$X_i \in 2^U$ | 1≤i≤|U|}으로서 다음의 조건을 만족한다.

(1)∀i, $X_i \neq \varnothing$

(2)∀i≠j, $X_i \cap X_j = \varnothing$

(3)$\bigcup \{X_i \mid 1 \le i \le n\} = U$

즉, (1) π(m, k)의 요소들(X_i:1≤i≤|U|)은 공집합이 아니다.

(2) π(m, k)의 요소들은 공통원소를 갖지 않는다.

(3) π(m, k)의 요소들을 합하면 객체의 전체집합 U가 된다.

표1과 같은 데이터 테이블의 속성 "키"에 대한 스케일 조합 트리(그림1)를 토대로 다음과 같이 4개의 파티션으로 구성되는 PS(키)를 추출할 수 있다.

PS(키) = {π(키, 0), π(키, 1), π(키, 2), π(키, 3)}

π(키, 0) = {{o1, o2, o3, o4, o5}},

π(키, 1) = {{o1, o2, o5}, {o3, o4}},

π(키, 2) ={{o2}, {o1, o5}, {o3}, {o4}},

π(키, 3) ={{o2}, {o1}, {o5}, {o3}, {o4}}.

[정의6] 주어진 데이터 테이블 S = (G, M, W, I)의 임의의 속성 m∈M에 대한 스케일 조합 트리 SCT(m)를 토대로 추출된 모든 파티션들의 집합 PS(m)이 주어졌을 때, 속성 m에 대한 입자개념계층구조 GCH(m)은 다음과 같이 정의한다.

GCH(m) = (V, E),

- $V = \bigcup_{i=0..|PS(m)|-1} \pi(\dot{m},i)$: 정점들의 집합(단, $\pi(\dot{m},i)$ ={i}×π(m,i))

- E ⊆ V×V : 변들의 집합

$$E = E \cup \{(x,y) \mid \forall y \in \pi(\dot{m},i+1) \exists x \in \pi(\dot{m},i) : y \subseteq x,$$
$$\text{단}, 0 \le i \le |PS(m)| - 2\}$$ ■

속성 m에 대한 입자개념계층구조 GCH(m)은 정점들의 집합 V와 변들의 집합 E로 구성된다. 정점들의 집합 V는 모든 파티션들의 원소들로 구성되며, 변들의 집합 E는 정점과 정점의 쌍으로 구성된다. 즉, 임의의 파티션 $\pi(\dot{m},i)$, $\pi(\dot{m},i+1)$ ∈PS(m)의

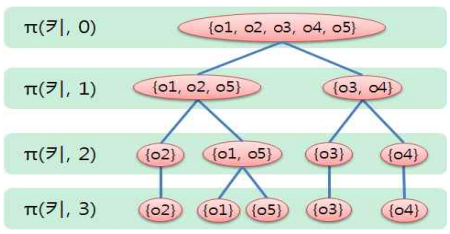

그림 2. PS(키)에 대한 입자개념계층구조
Fig. 2. Granular Concept Hierarchy for PS(height)

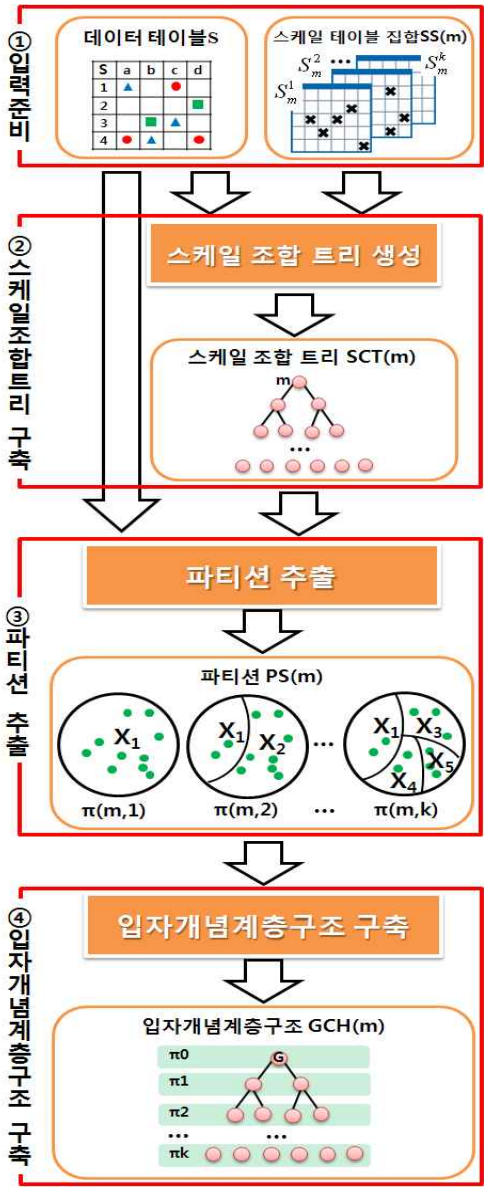

그림 3. 입자개념계층구조 구축을 위한 처리과정
Fig. 3. Process for constructing granular concept hierarchy

각 원소들 x, y($x\in\dot{\pi}(m,i)$, $y\in\dot{\pi}(m,i+1)$) 사이에 포함관계에 의해 x와 y를 연결하는 변을 E에 추가한다.

그림2는 그림1과 같은 스케일 조합 트리기반으로 추출된 파티션들의 집합 PS(키)를 토대로 구축된 입자개념계층구조이다.

각 노드들은 파티션의 원소를 나타내며, 각 변들은 원소들 사이에 부분집합 관계를 나타낸다. 입자개념계층구조는 스케일 조합 트리의 각 레벨에 따라 주어진 객체들을 다르게 분류하고 계층화한 것이다.

3.2 입자개념계층구조 구축을 위한 처리과정

본 절에서는, 3.1절에서 소개한 제반 정의를 토대로, 주어진 데이터 테이블의 속성에 대한 스케일 조합 트리 구축하고, 이를 토대로 파티션을 추출하여 입자개념계층구조를 구축하기 위한 처리과정을 정의한다.

그림 3은 주어진 데이터 테이블 S = (G, M, W, I)와 임의의 속성 m∈M에 대한 스케일 조합 트리를 구축하고 구축된 트리의 각 레벨에 대한 파티션들을 추출하여 입자화개념계층구조를 구축하기 위한 전체 처리과정을 나타낸 것이다.

① 입력준비

스케일 조합 트리 구축과 파티션 실행을 위한 준비 단계로써, 주어진 데이터 테이블 S와 S의 속성 m∈에 대한 스케일 테이블들의 집합 SS(m)을 입력한다.

② 스케일 조합 트리 구축

정의4를 토대로 만들어진 스케일 조합 트리 SCT(S, SS(m)) 알고리즘에 의해 입력된 데이터 테이블 S와 속성 m에 대한 스케일 테이블들의 집합 SS(m)으로부터 정보를 추출하여 스케일 조합트리를 생성한다.

SCT(S, SS(m))
//입력 : 데이터 테이블S=(G, M, W, I), 스케일테이블들의 집합
$SS(m)=\{S_m^1,S_m^2,...,S_m^k\}$ (단, $S_m^i=(W_m,\dot{M}_m^i,I_m^i)$)
//출력 : 스케일 조합트리 SCT(m)=(m:V, E)

```
begin
  V = VU{m};
  for i=1 to k
     V = VU(Ṁ_m^i ∪ W_m^i);
  end for
  for each x∈Ṁ_m^k
     E = EU{(m, x)};
  end for
  for i = 2 to k
     for each y∈Ṁ_m^(i-1), x∈Ṁ_m^i (s.t. y'⊆x')
        E = EU{(x, y)};
     end for
  for end
  for each (z, y) ∈ I_m^1
     E = EU{(y, z)};
  end for
end
```

③ 파티션 추출

주어진 데이터 테이블 S와 생성된 스케일 조합 트리를 입력하여, 정의5를 토대로 만들어진 파티션 Partition(S, SCT(m)) 알고리즘에 의해 입력된 스케일 조합 트리의 각 레벨에 대한 파티션들을 모두 추출한다. 즉, 스케일 조합 트리의 각 레벨에 따라 주어진 데이터 테이블 S의 객체들이 분류된다.

```
Partition(S, SCT(m))
//입력 : 데이터 테이블 S =(G, M, W, I), 스케일조합트리 SCT(m)=(m:V, E)
//출력 : 파티션들의 집합 PS(m)
begin
    for i = 0 to SCT(m).depth
        for each n∈V
            π(m,i) = π(m,i)∪ext(n);
        end for
        PS(m) = PS(m)∪π(m, i);
    end for
end
```

④ 입자개념계층구조 구축

파티션 추출단계(③)에서 추출된 파티션들 사이에 존재하는 상•하위관계를 토대로 입자개념계층구조를 구축한다. 즉, 스케일 조합 트리에 의해 분류된 객체들의 부분집합 관계를 토대로 계층구조를 형성한다.

```
GCH(PS(m))
//입력 : 파티션들의 집합 PS(m) = ∪_{i=0..k} π(m,i)
//출력 : 입자개념계층구조 GCH(m) = (V, E)
begin
    for i=0 to k=|PS(m)|
        V = V∪ π̇(m,i);
    end for
    for i = 0 to k
        for each x∈π(m, i), y∈π(m, i+1)(s.t. y⊆x)
            E = E∪{(x, y)};
        end for
    end for
end
```

IV. 지원도구의 개발 및 실험

본 장에서는, 제3장에서 제안한 기법을 지원하기 위해 개발한 지원도구를 소개한다. 또한, 본 논문에서 제안한 기법의 유용성을 검토하기 위해 개발된 도구를 사용하여 실제 데이터를 대상으로 실시한 실험에 대하여 설명한다.

4.1 지원도구의 개발

본 절에서는, 제3장의 제반 정의들과 알고리즘을 토대로, 다양한 관점과 추상화 수준을 고려하여 주어진 데이터를 분류하기위해 스케일 조합 트리를 기반으로 입자개념계층구조를 구축하기 위한 도구(G-Tool)를 개발하였다. G-Tool의 전체적인 아키텍처는 그림 4와 같이 3계층구조(UI Modules, Internal Data Set, Core Modules)로 구성되어 있다:

(1) Core Modules : Data Table Handler는 CSV 파일 형태로 표현된 데이터를 로드하여 데이터 테이블 형태로 표현한다. Scale Tables Handler는 입력된 데이터를 다양한 목적에 맞도록 적절히 분류하기 위해서 입력된 데이터의 각 속성들을 분석하고자하는 다양한 관점에 따라 각각의 스케일 테이블로 표현한다. Scale Combination Tree Constructor는 각 속성에 따라 입력된 스케일 테이블들 조합하여 스케일 조합 트리를 구축한다. Partition Extractor는 스케일 조합 트리를 토대로 입력된 데이터의 파티션들을 추출한다. Granular Concept Hierarchy Constructor는 추출된 파티션들 사이의 상·하위 관계를 파악하여 입자개념계층구조를 구축한다.

(2) Internal Data Set : Data Table은 [정의1]과 같이 여러 가지 다양한 값을 가지는 속성들과 객체들, 그리고 객체와 속성 사이의 관계를 나타낸다. Scale Tables은 [정의2]와 같이 주어진 데이터 테이블을 분석하고자 하는 목적 또는 관점에 따라서 이진 데이터 테이블로 표현되고 다양한 관점에서 정의되는 여러 가지 스케일 테이블들의 집합으로 구성된다. Scale Combination Tree는 [정의3]를 토대로 입력된 스케일 테이블들 사이의 상·하위 관계를 파악하여 [정의4]와 같은 스케일 조합 트리를 나타낸다. Partition은 [정의5]와 같이 Scale Combination Tree Constructor에 의한 구축된 스케일 조합 트리를 기준으로 각 레벨에 따라 입력된 데이터 테이블의 객체들을 분류하여 나타낸다. Granular Concept Hierarchy는 [정의6]과 같이 스케일 조합 트리의 각 레벨에 따라 주어진 데이터의 객체들을 다르게 분류하고 계층화한 것이다.

(3) UI Modules : Data Table & Scale Tables View는 CSV파일 형태로 입력된 데이터를 Data Table Handler 모듈을 적용하여 사용자에게 데이터 테이블 형태로 나타내고 데이터 테이블 생성 및 편집 기능도 제공하며, 또한, 입력 데이터를 분류하는 기준이 되는 스케일 테이블들을 이진

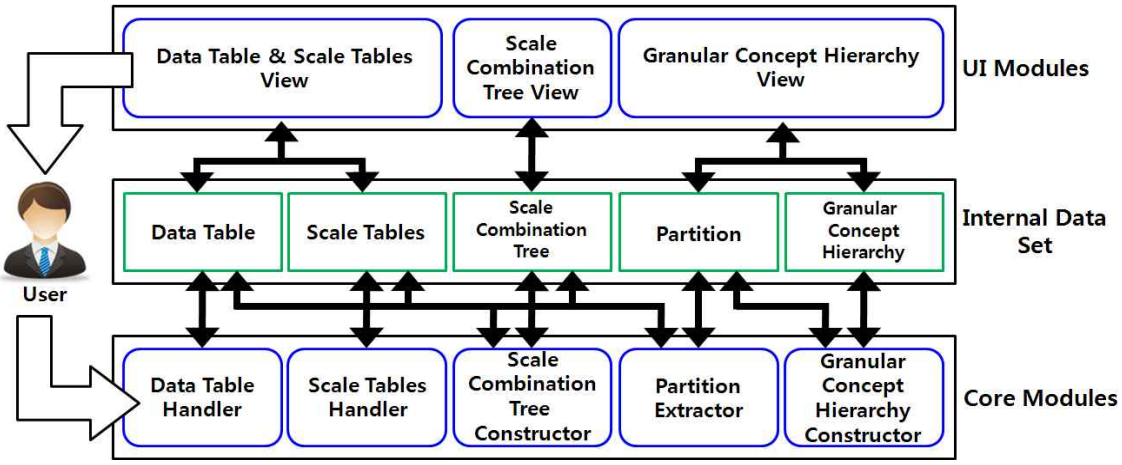

그림 4. G-Tool의 아키텍처
Fig. 4. Architecture of G-Tool

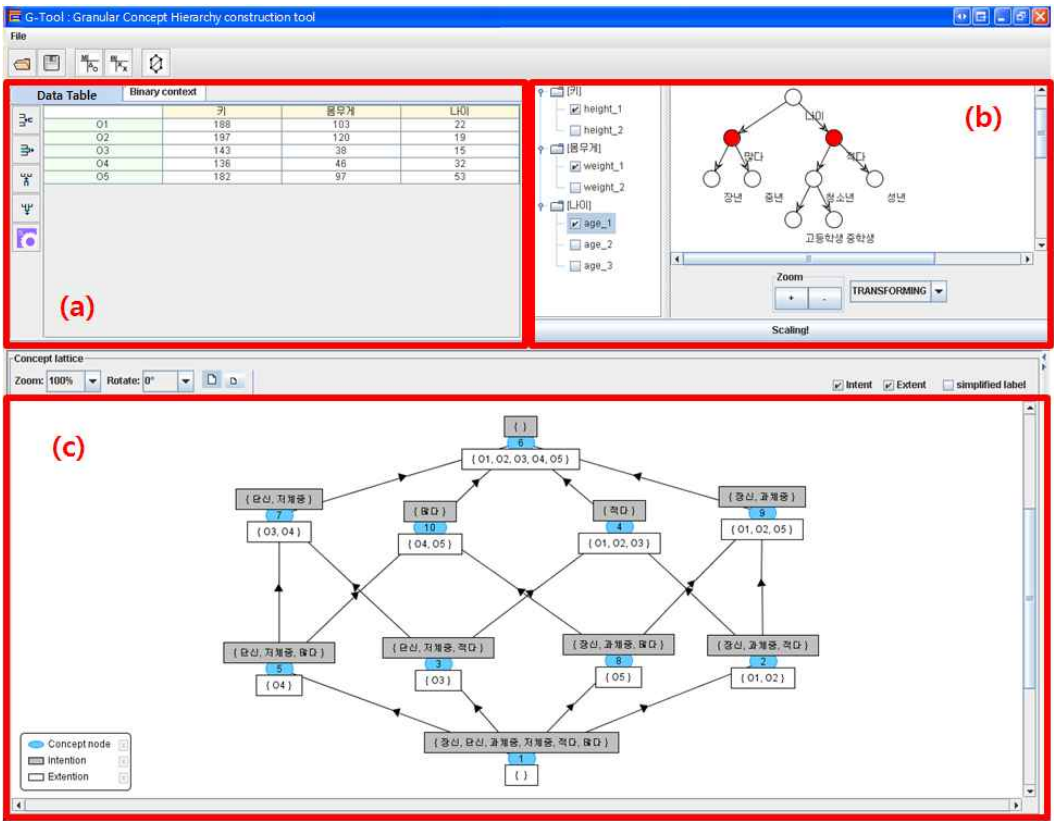

그림 5. G-Tool의 실행화면
Fig. 5. A screenshot of G-Tool

데이터 테이블 형태로 보여준다(그림5 (a)). Scale Combination Tree View는 Scale Combination Tree Constructor 모듈에 의해 구축된 스케일 조합트리를 가시화하고, 입력된 데이터를 분석하고자 하는 관점에 따라 스케일

정도를 선택한다(그림5 (b)). Granular Concept Hierarchy는 선택된 스케일 정도에 따라 입력된 데이터로부터 Partition Extractor 모듈에 의해 파티션을 추출하고 추출된 정보를 토대로 Granular Concept Hierarchy Constructor 모듈에 의해 입자개념계층구조를 구축하여 가시화한다(그림5 (c)).

4.2 실험

본 절에서는, UCI Machine Learning Repository[10]에 공개된 Car Evaluation Data Set을 대상으로 본 연구에서 제안한 기법을 적용하여 그 유용성을 검토한 실험의 결과를 요약하였다.

car	buying	maint	doors	persons	lug-boot	safety
c1	vhigh	vhigh	2	2	small	low
c2	vhigh	vhigh	2	2	small	med
c3	vhigh	vhigh	2	2	small	high
c4	vhigh	vhigh	2	2	med	low
c5	vhigh	vhigh	2	2	med	med
c6	vhigh	vhigh	2	2	med	high
c7	vhigh	vhigh	2	2	big	low
c8	vhigh	vhigh	2	2	big	med
c9	vhigh	vhigh	2	2	big	high
c10	vhigh	vhigh	2	4	small	low
				(중략)		
c1722	low	low	5more	more	small	high
c1723	low	low	5more	more	med	low
c1724	low	low	5more	more	med	med
c1725	low	low	5more	more	med	high
c1726	low	low	5more	more	big	low
c1727	low	low	5more	more	big	med
c1728	low	low	5more	more	big	high

그림 6. 자동차 평가 데이터 집합
Fig. 6. Car Evaluation Data Set

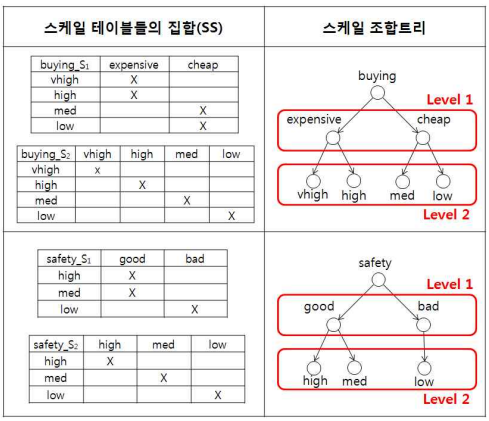

그림 7. 속성 "buying"과 "safety"에 대한 스케일 테이블들의 집합과 스케일 조합트리
Fig. 7. Set of scale tables and Scale Combination Trees for attributes "buying" and "safety"

Car Evaluation Data Set은 1728개의 자동차에 대한 6개의 속성(buying, maint, doors, persons, lug_boot, safety)과 각 속성들이 갖는 값들로 구성된 데이터 테이블이다(그림6 참조). Car Evaluation Data Set으로부터 임의로 20개의 자동차를 추출하고 본 연구에서 개발한 G-Tool을 사용하여 자동차의 가격과 안전성을 토대로 스케일 조합트리(그림 7)를 구성하고 스케일링 정도를 조절하면서 그림8 및 그림9와 같이 입자개념계층구조를 구축하였다.

그림8의 (a)는 속성 "buying"과 "safety"의 스케일 조합트리에서 Level 1을 기준으로 구성된 입자개념계층구조를 나타내고 있다. 자동차의 가격이 비싸면서 안전성이 좋은 자동차의 그룹(G6)과 가격은 비싸지만 안전성은 나쁜 자동차의 그룹(G7), 가격은 저렴하지만 안전성은 좋은 자동차의 그룹(G9)과 가격도 싸고 안전성도 나쁜 자동차의 그룹(G10)으로 입력 데이터를 분류하였다. 그림8의 (b)는 속성 "buying"의 스케일 조합트리에서 Level 1과 "safety"의 스케일 조합트리에서 Level 2를 기준으로 구성된 입자개념계층구조를 나타내고 있다. 그림9의 (b)에서, G7은 가격이 비싸면서 안전성은 가장 좋은 자동차의 그룹이고, G8은 가격이 비싸고 안전성은 중간정도인 자동차의 그룹, G9는 자동차는 비싸지만 안전성은 낮은 자동차의 그룹을 나타낸다. 또한, 자동차의 가격은 싸지만 안전성은 가장 좋은 자동차의 그룹은 G11이며, G12는 자동차의 가격은 싸지만 안전성은 중간정도의 자동차의 그룹을 나타내며, G13은 자동차의 가격도 싸고 안전성도 가장 낮은 자동차의 그룹이다. 그림9의 (a)는 속성 "buying"의 스케일 조합트리에서 Level 2와 "safety"의 스케일 조합트리에서 Level 1을 기준으로, 8개의 그룹(G6, G7, G9, G10, G12, G13, G15, G16)으로 분류된 입자개념계층구조를 나타내고 있으며, (b)는 속성 "buying"과 "safety"의 스케일 조합트리에서 Level 2를 토대로 11개의 그룹(G7, G8, G10, G11, G12, G14, G15, G16, G18, G19, G20)으로 구성된 입자개념계층구조를 나타내고 있다.

위의 실험결과(그림8, 그림9)에서 특히, 그림8의 (a)에 G9가 그림8의 (b)에서는 G11과 G12로 세분화되어 분류되었고, 그림9의 (a)에서도 G12와 G15로 분류되었음을 확인할 수 있다. 또한, 그림8의 (a)에 G9가 그림9의 (b)에서는 G14, G15, G18, G19로 그림8의 (b)와 그림9의 (a)에서보다 더 세분화되어 분류되었다. 따라서, 본 실험으로부터, 사용자의 의도 또는 목적에 맞추어서 스케일 조합트리를 구축하고 스케일링 정도를 조절하여 입자개념계층구조를 구축함으로써 사용자의 목적에 맞는 다양한 형태로 데이터를 분류할 수 있음을 확인하였다.

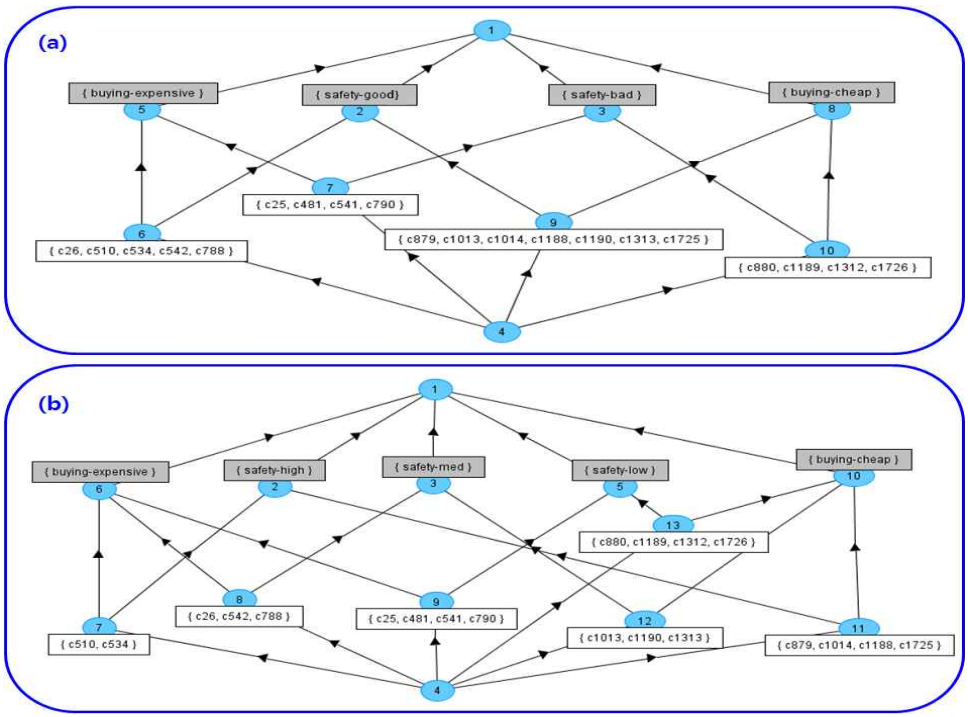

그림 8. 자동차의 가격과 안전성을 토대로 구축된 입자개념계층구조(1)
Fig. 8. Granular concept hierarchy constructed based on "buying" and "safety" of car(1)

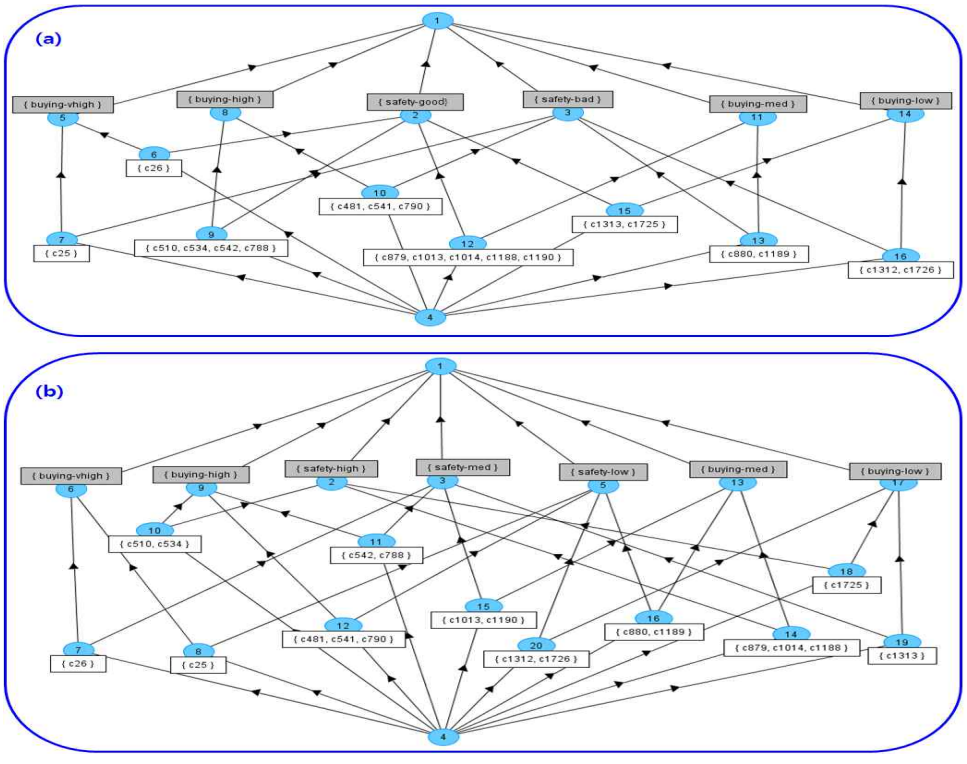

그림 9. 자동차의 가격과 안전성을 토대로 구축된 입자개념계층구조(2)
Fig. 9. Granular concept hierarchy constructed based on "buying" and "safety" of car(2)

V. 결 론

본 논문에서는, 주어진 데이터를 다양한 상황과 조건 및 추상화 수준을 적용하여 분류하기 위해 각 속성에 대한 스케일 조합트리를 구축하고 이를 토대로 스케일링 정도를 조절하여 다양한 유형의 입자개념계층구조를 구축하기 위한 방법을 제안하였다. 또한, 본 연구에서 제안한 기법을 지원하는 도구(G-Tool)를 개발하였으며, G-Tool을 사용하여 실제 데이터를 대상으로 본 연구에서 제안한 기법을 적용한 실험결과를 보고하였다. 본 연구에서 제안한 기법을 사용함으로써, 사용자의 의도 또는 목적에 맞추어서 스케일 조합트리를 구축하고 스케일링 정도를 조절하여 입자개념계층구조를 구축함으로써 사용자의 목적에 맞는 다양한 형태로 데이터를 분류할 수 있다. 또한, 구축된 입자개념계층구조를 토대로 속성들 사이의 연관규칙을 추출하여 사용자로부터의 질의에 대한 추론규칙을 생성할 수 있는 토대를 제공할 수 있다.

본 논문의 연구에서는, 데이터 분석 대상이 되는 객체집합으로부터 입자를 추출하기 위한 기준으로서 중복을 허용하지 않는 파티션(partition)을 이용하였으나, 보다 현실상황에 적합한 다양한 형태의 입자를 추출하기 위하여 중복을 허용하는 커버링(Covering)을 적용한 입자화 및 입자개념계층구조의 구축기법에 대한 후속연구가 진행될 예정이다.

참고문헌

[1] J.T. Yao, Y. Y. Yao and Y. Zhao, "Foundations of Classification", Foundations and Novel Approaches in Data Mining Studies in Computational Intelligence, Vol 9/2006, pp. 75-97, December, 2006.

[2] Wei-Zhi Wu, Yee Leung, Ju-Sheng Mi, "Granular Computing and Knowledge Reduction in Formal Contexts", IEEE Transactions on. Knowledge and Data Engineering, Vol. 21, Issue. 10, pp. 1461~1474, October, 2009.

[3] Ma Jian-Min, "Concept Granular Computing Systems", IEEE International Conference on. Fuzzy Systems and Knowledge Discovery, Vol. 1, pp. 150~154, December, 2009.

[4] B. Ganter, R. Wille, "Formal Concept Analysis: Mathematical Foundations", Springer, pp. 17~45, 1999.

[5] Guoyin Wang, Jun Hu, Qinghua Zhang, "Granular computing based data mining in the views of rough set and fuzzy set", IEEE Conference on. Granular Computing, August, 2008.

[6] S. Lee, J. Seo, "A Spatial Data Mining and Geographical Customer Relationship Management System", Journal of the Korea Society of Computer and Information v.15, n.6, pp. 121-128, 2010.

[7] G. Heo, J. Seo, I. Lee, "Problems in Fuzzy c-means and Its Possible Solutions", Journal of the Korea Society of Computer and Information v.16, n.1, pp.39-46, 2011.

[8] G. Stumme, "Hierarchies of Conceptual Scales", Workshop on Knowledge Acquisition, Modeling and Management (KAW '99), Vol. 2, pp. 78-95, Banff, Canada, October, 1999.

[9] R. Belohlavek, V. Sklenar, "Formal concept analysis over attributes with levels of granularity", International Conference on Computational Intelligence for Modelling, Control and Automation and International Conference on Intelligent Agents, Web Technologies and Internet Commerce, Vol. 1, pp. 619-624, Vienna, Austria, November, 2005.

[10] UCI Machine Learning Repository,http://archive.ics.uci.edu/ml/index.html

저 자 소 개

강 유 경

2003 : 선문대학교 컴퓨터정보학부 이학사.
2005 : 선문대학교 전자계산학과 이학석사.
2009 : 선문대학교 컴퓨터정보학과 이학박사
현 재 : 선문대학교 컴퓨터공학과 BK21 연구교수
관심분야 : Formal Concept Analysis, Semantic Web Mining, 소프트웨어공학 등
Email : yukyung.kang@gmail.com

황 석 형

1991 : 강원대학교 전자계산학과 이학사.
1993 : 일본 오사카대학교 정보공학과 공학석사.
1997 : 일본 오사카대학교 정보공학과 공학박사
2008 : 아일랜드국립대학교 DERI 객원연구원
현 재 : 선문대학교 컴퓨터공학과 교수
관심분야 : 소프트웨어공학, 객체지향, 온톨로지공학, 시맨틱 웹, Formal Concept Analysis, Semantic Web Mining
Email : shwang@sunmoon.ac.kr

김 응 희

2007 : 선문대학교 컴퓨터정보학부 이학사.
2009 : 서울대학교 치의과학과 공학석사.
현 재 : 서울대학교 의과대학 의료정보학과 박사과정
관심분야 : Data mining, Statistics
Email : eungheekim@snu.ac.kr

엄 태 정

2011 : 선문대학교 컴퓨터정보학과 이학사.
현 재 : 선문대학교 컴퓨터공학과 석사과정
관심분야 : 컴퓨터공학, 영상처리
Email : TJEom@mpcl.sunmoon.ac.kr