韓國컴퓨터情報學會 論文誌
第17卷 第1號, 2012. 1.

2012-17-1-1-3

구간데이터분석을 위한 형식개념분석기반의 분류

황 석 형*, 김 응 희**

A FCA-based Classification Approach for Analysis of Interval Data

Suk-Hyung Hwang*, Eung-Hee Kim**

요 약

다양한 정보기기와 소셜네트워크시스템, 그리고, 클라우드컴퓨팅환경 등과 같은 인터넷기반의 인프라를 토대로 분산화되고 공유가능한 데이터가 폭발적으로 증가하고 있다. 최근에는 데이터에 내재되어 있는 유용한 정보와 지식을 추출하고 분석 및 분류하기 위한 데이터분석 및 마이닝기법으로서, 이진데이터 또는 다치데이터에 관한 형식개념분석기법에 관한 연구가 활발하게 진행되어 다양한 분야에서 성공적으로 활용되고 있다. 그러나, 각 속성들이 구간값을 갖는 형태로 이루어진 구간데이터의 분석에 대한 형식개념분석에 관한 연구는 많이 수행되지 못하였다. 본 논문에서는, 구간데이터를 분석하기 위하여 형식개념분석기법을 기반으로 하는 새로운 분류기법을 제안한다. 또한, 구간데이터의 이진화, 개념추출 및 개념계층구조 구축 등, 본 논문에서 제안한 새로운 분류기법을 지원하기 위한 도구(iFCA)의 구축에 관하여 소개하고, 마지막으로, 몇가지 실세계의 데이터를 대상으로 한 실험결과를 토대로, 본 논문에서 제안하는 분류기법의 유용성에 대해서 설명한다.

▸ Keyword : 분류기법, 구간데이터, 형식개념분석, 이진화

Abstract

Based on the internet-based infrastructures such as various information devices, social network systems and cloud computing environments, distributed and sharable data are growing explosively. Recently, as a data analysis and mining technique for extracting, analyzing and classifying the inherent and useful knowledge and information, Formal Concept Analysis on binary or many-valued data has been successfully applied in many diverse fields. However, in formal concept analysis, there has been little research conducted on analyzing interval data whose attributes have

• 제1저자 : 황석형 • 교신저자 : 황석형
• 투고일 : 2011. 08. 26, 심사일 : 2011. 09. 29. , 게재확정일 : 2011. 10. 27.
* 선문대학교 컴퓨터공학과(Dept. of Computer Engineering, SunMoon University)
* 서울대학교 의생명지식공학연구실(Biomedical Knowledge Engineering Lab., Seoul National University)

some interval values. In this paper, we propose a new approach for classification of interval data based on the formal concept analysis. We present the development of a supporting tool(iFCA) that provides the proposed approach for the binarization of interval data table, concept extraction and construction of concept hierarchies. Finally, with some experiments over real-world data sets, we demonstrate that our approach provides some useful and effective ways for analyzing and mining interval data.

▸ Keyword : Classification, Interval Data, Formal Concept Analysis, Binarization

I. 서 론

최근 우수한 성능의 컴퓨터와 인터넷 기반의 정보인프라를 토대로, 다양한 모바일 정보기기 및 소셜네트워크시스템의 발달과 클라우드컴퓨팅환경의 등장에 의해 분산·공유가능한 데이터양이 폭발적으로 증가하고 있다. 특히, 거대한 데이터의 실시간 분석이 가능해지고 방대한 데이터로부터 다양한 분석을 통해 새로운 지식의 발견이 가능해짐에 따라서, 21세기 지식정보화 사회에서는 새로운 지식의 습득이 경쟁력의 원천이 되고 있고, 이를 뒷받침할 수 있는 다양한 기술들에 대한 관심이 높아지고 있다. 데이터마이닝(Data Mining)은, 방대한 양의 데이터에 함축적으로 내재되어 있는 관계, 패턴, 규칙 등을 찾아내서 모델화함으로써 유용하고 의미있는 정보들을 추출하기 위한 연구분야로서[1], 주로, 분류(Classification), 연관법칙추출(Association Rule Mining), 그리고 군집화(Clustering) 등과 같은 제반기법들에 관한 연구가 활발하게 진행되어 다양한 분야에서 활용되고 있다[2,3,4,5,6].

분류(Classification)기법의 일종으로서, 형식개념분석기법(Formal Concept Analysis)은 Ganter와 Wille[7]에 의해 창안되어, 데이터마이닝 분야 뿐만아니라, 온톨로지공학, 시맨틱 웹, 소프트웨어공학, 정보 검색, 그리고 의학 및 바이오인포메틱스 등, 다양한 분야에서 사용되고 있다[8,9]. 형식개념분석기법은 주어진 문제영역의 각 객체들(Objects)과 그들이 갖는 속성들(Attributes)로부터, 양쪽 집합의 요소들 사이에 내재되어 있는 갈루아 대응관계(Galois Connection)를 토대로 개념(Concept)이라는 지식기본단위로 추출하고, 개념들사이의 순서관계를 토대로 개념격자(Concept Lattice)를 구축함으로써 데이터의 분류와 군집화, 연관법칙의 추출 등과 같은 지식추출 및 표현, 그리고 추론을 위한 수학적인 모델을 제공해 준다.

형식개념분석기법에서는 실세계의 분석대상영역내의 객체들에 대해서, 각 객체들이 어떤 속성[1]을 소유하고 있는지 여부를 객체와 속성 사이의 이항관계로 파악하여 이진데이터테이블 형태로 입력된다. 그러나, 실세계의 분석대상 객체들은 대부분 이진속성만을 갖는 경우보다는 보다 복잡하고 다종다양한 속성값을 갖는 경우가 많다. 이와같이 분석대상객체가 다종다양한 값을 갖는 속성[2]을 소유하는 경우에는 변환처리과정, 즉, 이진화 과정("Scaling" 또는 "Binarization"이라고 부름)을 수행하여 이진데이터테이블 형태로 변환된다. 이와같은 이진화 과정에서는, 스케일(Scale)이라는 변환기준을 토대로 각 다치속성을 이진속성으로 변환하는 처리가 수행된다[7].

스케일은 분석대상영역의 지식을 토대로 제공되며, 형식개념분석기법의 사용자에 의해 특정한 스케일이 선택되어 사용된다. 또한, 스케일은 다치속성을 갖는 데이터에 대한 분석기준이 되므로 형식개념분석기법을 적용하는 경우, 중요한 부가적 정보로서 이용된다. 즉, 다치속성을 갖는 데이터에 대한 이진화과정을 수행하는 과정에서 어떠한 유형의 스케일을 기준으로 설정하였는가에 따라서 상이한 이진테이터테이블이 구성되므로, 결과적으로 이진화과정에서 사용되는 스케일에 따라서 형식개념분석에서는 동일한 데이터에 대해서 다양한 분석결과를 얻을 수 있다. 따라서, 형식개념분석기법에서는 다양한 분석결과를 얻기 위해서는 주어진 데이터의 특성에 알맞은 다양한 해석기준(즉, 스케일)을 제공할 필요가 있다. 분석대상영역의 여러가지 상황과 지식을 토대로 형식개념분석을 수행하기 위한 대표적인 이진화기법으로는, 개념형 이진화기법(Conceptual Scaling)[10]과 논리형 이진화기법(Logical Scaling)[11], 그리고 관계형 이진화기법(Relational Scaling)[12] 등이 제안되어 다양한 유형의 데이터분석에 적용된 사례들이 보고되고 있다. 그러나, 현재까지 형식개념분석기법에서는 주로 이진데이터 및 다치데이터

1) 이때, 각 객체가 소유하고 있는 속성을 "이진속성(binary attribute)" 또는 "One-valued attribute"라고 부른다.

2) "다치속성(多値屬性)" 또는 "many-valued attribute"라고 부른다.

에 한정되어 있고, [최소값, 최대값]으로 구성되는 구간데이터(interval data)에 대해서는 고려하지 못하였다.

따라서, 본 논문에서는, 구간데이터를 다치속성의 값으로 갖는 데이터에 대한 데이터마이닝기법의 일종으로서, 형식개념기법을 기반으로 하는 분류기법을 제안한다. 구체적으로는, 구간데이터를 속성값으로 갖는 객체들에 대하여, 객체들 사이의 유사관계를 토대로, 형식개념분석기법을 확장·정의하여 구간값을 다치속성의 값으로 갖는 데이터에 적용가능한 새로운 분류기법을 제안한다. 새로운 분류기법을 지원하기 위한 도구의 구축과 더불어서, 몇가지 실세계의 데이터를 토대로 실험을 수행하여, 본 논문에서 제안하는 분류기법의 유용성을 확인하고, 향후 연구과제에 대해서 설명한다.

본 논문의 구성은 다음과같다. 제2장에서는 형식개념분석기법 및 본 논문에서 필요한 제반 정의들과 개념들에 대해서 소개하고, 제3장에서는 구간값을 다치속성값으로 갖는 객체들 사이의 유사관계를 토대로 형식개념분석기반의 분류기법을 정의한다. 제4장에서는 새로운 분류기법을 지원하는 도구와 실세계의 데이터를 토대로 수행한 실험결과를 보고하고, 제5장에서는 결론과 향후 연구과제에 대해서 설명한다.

II. 형식개념분석기법 및 제반정의들

형식개념분석기법에서는 분석대상이 되는 데이터를 표1과 같은 데이터테이블(formal context라고 부르는 일종의 이진 데이터테이블) 형태로 입력받는다. 즉, 데이터테이블 K=(G, M, I)는 분석대상이 되는 데이터를 구성하는 객체들의 집합 G와 속성들의 집합 M, 그리고 G와 M사이의 이항관계 I⊆G×M로 이루어진다. 즉, 임의의 객체g∈G와 속성m∈M에 대하여, (g, m)∈I(또는 gIm)은 객체 g가 속성 m을 가지고 있다는 것을 의미하며, 해당 셀에 관련된 객체와 속성이 이항관계 I를 만족할 경우에는 ×표시하고, 이외의 경우에는 빈 공란으로 남겨둔다[7].

표 1. 이진데이터테이블
Table 1. Binary Data Table

	a1	a2	a3	a4	a5
o1			×	×	
o2	×	×	×		×
o3	×			×	×
o4			×		

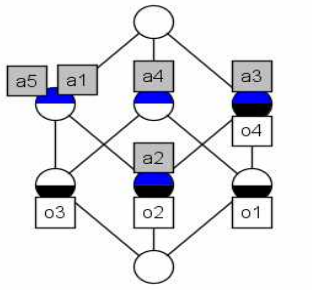


그림 1. 표1의 데이터테이블에 대한 개념격자
Fig. 1. Concept Lattice for Data Table 1

데이터테이블 K=(G, M, I)의 임의의 O⊆G, A⊆M에 대하여, intent(O)=A ∧ extent(A)=O를 만족하는 (O, A)를 개념(formal concept)이라고 부른다(단, intent(O) := {a∈M|∀o∈O:(o, a)∈I}, extent(A) := {o∈G|∀a∈A:(o, a)∈I}이며, 개념(O, A)를 구성하는 객체집합O를 외연(extension), 속성집합A를 내포(intention)라고 부른다). 즉, 개념 (O, A)는 객체집합O의 모든 객체들이 공통적으로 갖는 속성들의 집합이 A와 같고, 속성집합A의 모든 속성들을 공통적으로 갖는 객체들의 집합이 O와 같음을 의미한다. 예를들면, 표1의 데이터테이블로부터 ({o2, o3}, {a1, a5}), ({o2}, {a1, a2, a3, a5}) 등과 같은 개념들을 추출할 수 있다. 특히, 임의의 개념 (O1, A1)와 (O2, A2)에 대하여, O1⊆O2(⇔A1⊇A2)라면, (O1, A1)은 (O2, A2)의 하위개념(또는, (O2, A2)는 (O1, A1)의 상위개념)이라고 부르며, (O1, A1)≤(O2, A2)와 같이 표기한다. 예를들면, ({o2}, {a1, a2, a3, a5})는 ({o2, o3}, {a1, a5})의 하위개념이다.

데이터테이블K로부터 추출된 모든 개념들로 구성된 집합 B(K):={(O, A)| intent(O)=A ∧ extent(A)=O, O⊆G, A⊆M}와 개념들 사이의 상하위관계≤로 이루어진 계층구조 B(K):=(B(K),≤)를 개념격자(Concept Lattice)라고 부른다. 개념격자는 그림1과 같은 Line Diagram으로 표현할 수 있다. Line Diagram에서는, 각 개념들과 이들 사이의 상하위관계가 링크에 의해 표시되며, 특히, 개념들 간의 링크에 의해 만들어지는 경로에 의해 상위개념으로부터 하위개념으로 속성들이 상속되며, 하위개념으로부터 상위개념으로 해당 객체들이 전파된다. 주어진 이진데이터테이블로부터 개념을 추출하고 개념격자형태로 계층화하여 표현함으로써, 문제영역내에 내재되어있는 지식을 개념단위로 추출하여 분류하고 체계화할 수 있다.

표 2. 다치데이터테이블
Table 2. Many-Valued Data Table

	m1	m2
g1	0	21
g2	3	50
g3	6	66
g4	6	88
g5	9	17

표 3. Nominal scale의 예
Table 3. An Example of Nominal scale

Sm1	m1=0	m1=3	m1=6	m1=9
0	×			
3		×		
6			×	
9				×

표 4. Biordinal scale의 예
Table 4. An Example of Biordinal scale

Sm2	m2<18	m2<40	m2<=65	m2>65	m2>=80
17	×	×	×		
21		×	×		
50			×		
66				×	
88				×	×

한편, 실세계의 분석대상 데이터들은 객체가 어떤 속성을 소유하는지 여부를 이진값으로 나타내는 이진속성값을 갖는 경우보다는 보다 복잡하고 다종다양한 속성값을 갖는 경우가 많다. 형식개념분석기법은 다치속성을 갖는 데이터에 대해서도 적용가능하며, 다치속성을 포함하는 데이터테이블을 다치데이터테이블(many-valued context 또는 many-valued data table)이라고 부른다[7-9].

[정의 1] 다치데이터테이블 K=(G, M, (Wm)m∈M, I)는 객체들의 집합 G와 다치속성들의 집합 M, 속성값의 집합 (Wm)m∈M, 그리고 객체와 속성, 속성값들사이의 삼항관계 I⊆{(g, m, w)|g∈G m∈M, w∈Wm}로 구성된다.

즉, G와 M, 그리고 Wm의 원소들은 각각 해당 데이터테이블의 객체들과 각 객체들이 가질 수 있는 속성들, 그리고 해당 속성의 값들을 나타낸다. (g, m, w) ∈ I 또는 m(g)=w는, 객체 g가 속성 m을 가지고 있고 해당 속성의 값이 w임을 의미한다. 다치데이터테이블은 표2와같은 형태로 나타낼 수 있으며, 각 셀에는 해당 객체가 갖는 속성의 값을 표시한다.

표 5. 표3과 표4의 변환규칙을 토대로 표2의 다치데이터테이블을 이진화하여 얻어진 이진데이터테이블
Table 5. Scaled Binary Data Table of Table 2 based on Table 3 and 4

	m1 =0	m1 =3	m1 =6	m1 =9	m2 <18	m2 <40	m2 <=65	m2 >65	m2 >=80
g1	×					×	×		
g2		×					×		
g3			×					×	
g4			×					×	×
g5				×	×	×			

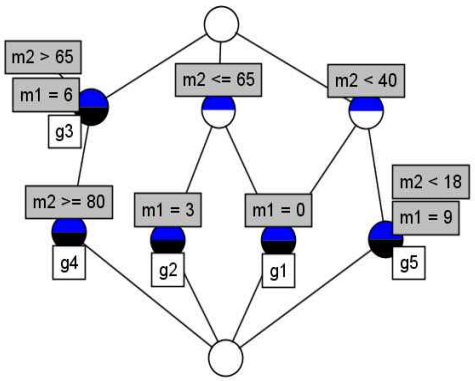


그림 2. 표5의 개념격자
Fig.2. Concept Lattice of Table 5

이와 같은 다치데이터테이블로부터 개념들을 추출하고 개념격자를 구성하기 위해서는 개념을 추출하기에 앞서서, 주어진 변환규칙에 따라 다치데이터테이블을 이진데이터테이블로 변환하기 위한 이진화과정(Binarization)이 필요하다. 이진화과정에서는 다양한 변환규칙들을 토대로 각 다치속성들을 이진속성으로 변환시켜서 새로운 이진테이터테이블을 구성한다. 이러한 이진화과정에서 사용되는 변환규칙(scale)은 다음과 같이 정의된다.

[정의 2] 다치데이터테이블 K=(G, M, (Wm)m∈M, I)의 임의의 속성 m∈M에 대한 개념분석을 위한 변환규칙 Sm은 다음과 같은 이진데이터테이블로 정의된다:

Sm:=(Wm, Mm, Im).

이진화과정에서 사용되는 변환규칙들은 문제영역의 전문지식에 따라서 다양하게 정의되며, 형식개념분석을 수행하는 목적에 따라서 사용자가 선택하게 된다. 일반적으로, 각 변환규칙들은 Nominal Scale, Biordinal Scale, Ordinal Scale, 등과 같이 유형화할 수 있다[7, 8].

[정의 3] 주어진 임의의 m∈M에 대한 변환규칙 Sm:=(Wm, Mm, Im)을 다치데이터테이블 K=(G, M, (Wm)m∈M, I)에 적용하여 이진화과정을 수행하면 다음과 같은 이진데이터테이블D(K)로 변환된다. 즉, D(K) = (G, ∪m∈M{m}×Mm, J), 이때, J⊆G×(∪m∈M{m}×Mm)이고 gJ(m, w)⇔(g, m, w)∈I이다.

예를 들어, 표2의 다치데이터테이블의 속성 m1과 m2에 대해서는, 표3과 표4의 변환규칙들을 각각 적용하여 표5와 같은 이진데이터테이블로 변환될 수 있으며, 이를 토대로 그림2와 같은 개념격자를 추출할 수 있다.

III. 형식개념분석을 기반으로하는 구간데이터의 분류기법

3.1 구간값을 속성의 값으로 갖는 객체들사이의 유사관계

실세계의 분석대상 데이터들중에는 속성값으로서 표6과 같이 구간값, 즉, [최소값, 최대값]과 같은 형태의 값을 갖는 경우를 흔히 볼 수 있다. 예를들면, 참고문헌[13, 14]에서는 33대의 자동차에 대하여 가격, 엔진성능, 최고속력 등의 7개 속성들에 대하여 구간값이 주어져 있다(그림8참조). 이와같은 형태의 데이터집합은, 각 객체g∈G의 속성m∈M에 대하여 구간값을 갖는 형태(즉, m(g)=[x, y])의 데이터테이블, 즉, 구간데이터테이블(interval-valued data table)이라고 부르며, 다음과 같이 정의한다.

[정의 4] 구간데이터테이블 K=(G, M, [Wm]m∈M, I)는 다음과 같은 요소들을 갖는 데이터테이블이다 :

- G : 객체들의 집합
- M : 속성들의 집합
- [Wm]m∈M=∪m∈M{[a, b]|∃g∈G:m(g)=[a, b]}
  : 각 속성들에 대한 구간값들의 집합
- I⊆{(g, m, [a, b])|g∈G, m∈M, [a, b]∈[Wm]m∈M}
  : 객체, 속성, 구간속성값들 사이의 삼항관계.

표6의 구간데이터테이블에서, m1(g1)=[1,3], m1(g2)=[2, 5], m1(g3)=[6, 7] 등 이다. 특히, 구간값의 최대값 및 최소값이 같은 경우, 즉,m(g)=[x, x]인 경우, m(g)=x와 같이 나타낼 수 있다. 예를들어, 표8에서 m1(g5)=[6, 6]=6이다. 따라서, 다치데이터테이블은 구간데이터테이블의 일종으로 볼 수 있다.

표 6. 구간데이터테이블
Table 6. Interval Data Table

	m1	m2	m3
g1	[1, 3]	[8, 12]	[1, 2]
g2	[2, 5]	[13, 18]	[2, 5]
g3	[6, 7]	[8, 11]	[4, 5]
g4	[2, 3]	[9, 18]	[6, 9]
g5	[6, 6]	[8, 11]	[7, 10]

주어진 데이터를 분석하기 위한 일반적인 방법으로서, 어떤 기준을 토대로 데이터들을 그룹화하여 공통속성을 파악하는, 유사관계기반의 분류기법이 많이 사용된다[4,5]. 특히, 형식개념분석기법[7]에서는 다치데이터테이블K의 임의의 속성m에 대하여, 주어진 임계값Θ에 관한 두 객체g1, g2사이의 유사관계 $\cong_\theta^m$는, 각 객체가 갖는 속성값의 차를 고려하여, $g1 \cong_\theta^m g2 \Leftrightarrow |m(g1)-m(g2)| \le \Theta$와 같이 정의할 수 있고, 구간데이터테이블의 객체g1, g2(단, m(g1)=[a,b], m(g2)=[c,d])의 경우, 두 객체g1, g2사이의 유사관계는 다음과 같이 정의할 수 있다.

[정의 5]구간데이터테이블 K=(G, M, [Wm]m∈M, I)에서 구간값을 갖는 임의의 속성m∈M에 대하여, 주어진 임계값Θ에 관하여 두 객체g1, g2∈G(단, m(g1)=[a, b], m(g2)=[c, d])사이에 다음과 같은 조건이 만족될 때, 두 객체g1, g2는 유사관계에 있다라고 하고, $g1 \cong_\theta^m g2$로 나타낸다. 즉,

$$g1 \cong_\theta^m g2 \Leftrightarrow \exists [a, b]=m(g1),\ [c, d]=m(g2):$$
$$\max(b, d)-\min(a, c) \le \Theta.$$

유사관계 $\cong_\theta^m$는, 반사성과 대칭성은 만족하지만, 추이성은 만족하지 않는다.

임의의 데이터테이블에는, 다종다양한 값을 갖는 여러 개의 속성들(이진속성, 다치속성, 구간속성 등)이 포함될 수 있으므로, 형식개념분석기법을 적용하기 위해서는 각 객체들이 소유하는 각 속성들의 다양한 특징들을 충분히 고려하여 객체들 사이의 유사관계를 파악하여 적합한 개념을 추출할 수 있어야 한다. 특히, 구간데이터테이블에서 모든 객체들 사이의 유사관계는 각 속성에 대하여 주어진 구간값들 사이의 유사관계를 모두 파악해야하며, 이러한 관계들은 구간데이터테이블의 이진화과정에서 변환규칙으로 이용할 수 있다.

[정의 6]주어진 구간데이터테이블 K=(G, M, [Wm]m∈M, I)의 임의의 속성m∈M과 임계값Θ에 대하여, 구간값의 집합 Wm과 객체들 사이의 유사관계 $\cong_{\theta}^{m}$ 를 추출하여 작성된 이진데이터테이블TK(m, Θ)은 다음과 같이 정의된다.

$$TK(m, \Theta) = (Wm, Wm, \cong_{\theta}^{m}).$$

예를들어, 표6의 속성m3과 임계값Θ=5가 주어졌을 때, TK(m3, 5)와 이에 대응하는 개념격자는 각각 표7 및 그림3과 같다.

유사관계 $\cong_{5}^{m3}$ 는 반사성과 대칭성을 만족하므로, 표7의 TK(m3, 5) 또한 반사성과 대칭성을 갖는다. 또한, TK(m3, 5)로부터 생성된 개념격자(그림 3)는 대칭적인 구조를 갖는다. 예를들면, 그림3의 개념격자에서 좌측상단에 있는 개념({[1,2], [2,5], [4,5]}, {[4,5]})은 좌측하단에 있는 개념({[4,5]}, {[1,2], [2,5], [4,5]})과 대응한다. 이와같이 대칭적인 구조를 갖는 개념격자에서는 외연과 내포가 동일한 개념들을 대칭축으로하여 각 개념들사이의 대칭관계가 형성된다. 즉, 주어진 구간데이터테이블 K=(G, M, [Wm]m∈M, I)의 임의의 속성m∈M에 대하여, 유사관계 $\cong_{\theta}^{m}$ 를 추출하여 작성된 이진데이터테이블 TK(m, Θ) = (Wm, Wm, $\cong_{\theta}^{m}$)로부터 생성된 개념격자 B(TK(m, Θ))에서는 다음과 같이 대칭축에 존재하는 개념들과 상위 및 하위에 존재하는 개념들의 집합 등, 3종류의 개념집합들로 분류할 수 있다.

[정의7]이진데이터테이블TK(m,Θ) = (Wm, Wm, $\cong_{\theta}^{m}$)로부터 구축된 개념격자B(TK(m, Θ))에서, 개념들의 집합B(TK(m, Θ))의 원소들(개념들)은 다음과 같은 3종류의 개념집합들로 분류된다. 즉, B(TK(m,Θ)) = S(TK(m,Θ)) ∪ U(TK(m,Θ)) ∪ L(TK(m,Θ)), 단, S(TK(m,Θ))={(A,B)∈B(TK(m,Θ))|A=B}는 대칭축의 개념집합, U(TK(m,Θ))={(A,B)∈B(TK(m,Θ))|(A,B)≥(X,Y)∈S(TK(m,Θ))}는 상위의 개념집합,L(TK(m,Θ))={(A,B)∈B(TK(m,Θ))|(A,B)≤(X,Y)∈S(TK(m,Θ))}는 하위의 개념집합.

그림3으로부터, 다음과 같은 개념집합들을 구할 수 있다.

S(TK(m3,5))={({[1,2], [2,5]}, {[1,2], [2,5]}), ({[4,5], [6,9]}, {[4,5], [6,9]}),({[6,9], [7,10]}, {[6,9], [7,10]})},

U(TK(m3,5))={({[1,2], [2,5], [4,5]}, {[4,5]}), ({[4,5], [6,9], [7,10]}, {[6,9]})} ∪ S(TK(m3,5)),

L(TK(m3,5))={({[4,5]}, {[1,2], [2,5], [4,5]}), ({[6,9]}, {[4,5], [6,9], [7,10]})} ∪ S(TK(m3,5)).

표 7. 이진데이터테이블 TK(m3, 5)
Table 7. Binary Data Table TK(m3, 5)

	[1,2]	[2,5]	[4,5]	[6,9]	[7,10]
[1,2]	×	×	×		
[2,5]	×	×	×		
[4,5]	×	×	×	×	
[6,9]			×	×	×
[7,10]				×	×

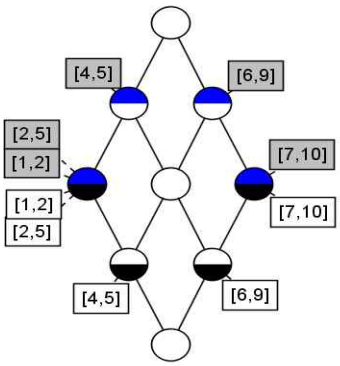


그림 3. 개념격자 B(TK(m3, 5))
Fig. 3. Concept Lattice B(TK(m3, 5))

[정의 8]구간데이터테이블 K=(G, M, [Wm]m∈M, I)에 대하여, 임의의 속성m∈M과 주어진 임계값Θ에 관한 유사클래스집합 CS(m, Θ)는 다음과 같이 정의된다.

CS(m, Θ):=∪c∈U(TK(m,Θ)){[x,y]|∀[a,b],[c,d]∈intent(c):

$$\min(a, c)=x \wedge \max(b, d)=y\}$$

예를들어, 속성m1, m2, m3의 임계값이 각각 3, 9, 5로 주어졌을 때, 각 속성에 대한 유사클래스들의 집합은 다음과 같다. 즉,

CS(m1, 3)={[2, 3], [1, 3], [2, 5], [6, 7]},

CS(m2, 9)={[8, 12], [9, 18]},

CS(m3, 5)={[4, 5], [6, 9], [1, 5], [4, 9], [6, 10]}.

주어진 구간데이터테이블 K=(G, M, [Wm]m∈M, I)에 대하여, 각 속성m∈M에 대한 주어진 각 임계값Θ에 관한 유사관계 $\cong_{\theta}^{m}$ 를 기반으로 유사클래스집합CS(m, Θ)을 구하고, 이를 토대로 이진화과정을 수행함으로써, 다음과 같은 이진데이터테이블 D(K) = (G′, N, J)로 변환할 수 있다. 단, G′ = G,

$N = \cup m \in M(\{m\} \times CS(m, \Theta))$, $J \subseteq (G' \times N)$:

$(g, (m, CS(m, \Theta))) \in J$

$\Leftrightarrow \exists c \in U(TK(m, \Theta))[\forall [a, b], [c, d] \in intent(c):$

$x=min(a, c) \wedge y=max(b, d)].$

3.2 구간데이터에 대한 유사관계를 기반으로 하는 형식개념분석기법

구간값을 갖는 데이터테이블이 주어졌을 때, 유사관계를 기반으로 이진데이터테이블로 변환하고, 개념을 추출하여 개념격자를 구축하는 등, 일련의 작업을 수행함으로써 형식개념분석기법을 기반으로하는 구간데이터의 분류작업을 수행할 수 있다. 이상에서 논의한 일련의 분석과정은 3.2절의 제반 정의들을 토대로 다음과 같이, 구간데이터에 대한 유사관계 $\cong_\theta^m$ 를 기반으로하는 형식개념분석을 위한 알고리즘 iFCA(K)로 정리할 수 있다.

iFCA(K)

- 입력 : K=(G, M, [Wm]m∈M, I), {Θm | m∈M}
- 출력 : 개념격자 B(D(K))

1. 구간데이터테이블 K로부터 유사관계 $\cong_\theta^m$ 를 기반으로하는 변환규칙TK(m, Θ) 추출
   for each m∈M do
      TK(m, Θ) = (Wm, Wm, $\cong_\theta^m$);
   end_for
2. 유사관계 $\cong_\theta^m$ 를 기반으로하는 구간데이터테이블의 이진화과정 수행
   for all m∈M do
      D(Km) = Binarization(K, m, Θm, TK(m, Θ));
   end_for
3. 각 속성들에 대한 이진화과정 수행결과들(|M|개의 이진데이터테이블들)을 병합
   for all D(K1), D(K2)∈{D(Km)|Km=(G, m, [Wm], I∩(G×{m})) for all m∈M} do
      Let D(K1) := (G, M1, I1), D(K2) :=(G, M2, I2);
      D(K)=$\left(G, \dot{M}_1 \cup \dot{M}_2, \dot{I}_1 \cup \dot{I}_2\right)$,
      단, $\dot{M}_i := \{i\} \times M_i$, $\dot{I}_i := \{i\} \times I_i$ for $i \in \{1, 2\}$;
   end_for
4. 이진데이터테이블로부터 개념추출
   B(D(K)) = ExtractConcepts(D(K));
5. 개념격자 구축
   B(D(K)) = BuildConceptLattice(B(D(K)));
Return B(D(K));

Function Binarization(K, m, Θ, TK(m, Θ))

- 입력 : 구간데이터테이블 K=(G, M,[Wm]m∈M, I), m, Θ
- 출력 : 이진데이터테이블 D(K)=(G', N, J)

1. ExtractConcepts(TK(m, Θ));
2. S(TK(m, Θ))={(A, B)∈B(TK(m, Θ))|A=B};
3. U(TK(m,Θ))={(A,B)∈B(TK(m,Θ))|(A,B)≥(X,Y)∈S(TK(m,Θ))};
4. for all c∈U(TK(m, Θ)) do
      CS(m,Θ)=CS(m,Θ)∪{[x,y]|∀[a,b],[c,d]∈intent(c):min(a,c)=x∧max(b,d)=y};
   end_for;
5. G' = G
6. N = ∪m∈M{m}×CS(m, Θ));
7. for all g∈G' do
      if m(g)∈CS(m,Θ) then J=J∪{(g, (m, CS(m,Θ)))};
   end_for;
8. Return (G', N, J);

Function ExtractConcepts(K)

- 입력 : 이진데이터테이블 K=(G, M, I)
- 출력 : 개념집합 B(K)

1. for all g∈G do
2. B(K) = B(K)∪{ (extent(intent(g), intent(g)) };
3. end_for
4. for all c∈B(K) do
5. for all g∈(G-extent(c)) do
6. X = extent(c)∪{g};
7. if (extent(intent(X)), intent(X))∉B(K)) then
8. B(K)=B(K)∪{(extent(intent(X)), intent(X))};
9. end_if
10. end_for
11. end_for
12. Return B(K);

Function BuildConceptLattice(B(K))

- 입력 : 개념집합 B(K)
- 출력 : 개념격자 B(K) = (B(K), E≤)

1. for all c1∈B(K) do
2. for all c2∈B(K)-{c1} do
3. if ((c1≤c2)∧∄c3∈(B(K)-{c1, c2})[
4. c1≤c3 ∧ c3≤c2]) then
5. E≤ = E≤ ∪ {(c1, c2)};
6. end_for
7. end_for
8. Return B(K);

표6의 구간데이터테이블에 대하여, iFCA를 적용하면, 표8과 같은 이진데이터테이블로 변환할 수 있으며, 그림4와 같은 개념격자들을 추출할 수 있다. 그림4의 개념격자는, 객체 g1~g5가 각 속성m1, m2, m3 및 해당 임계값Θmi(i=1, 2, 3)에 대하여 유사관계 $\cong_\theta^m$ 를 기반으로 어떻게 분류되는지를 나타낸다.

표 8. 이진화과정에 의해 변환된 이진데이터테이블
Table 8. Binary Data Table after Binarization

g1	×				×		×				
g2			×			×	×				
g3				×	×		×	×	×		
g4	×	×	×			×			×	×	×
g5				×	×						×

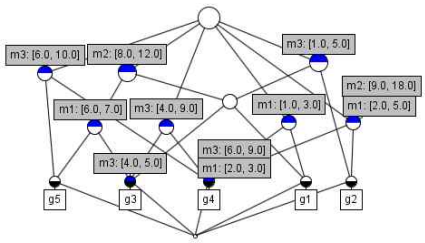
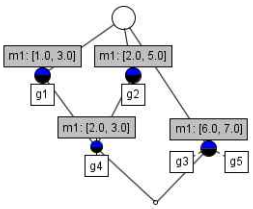
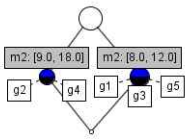


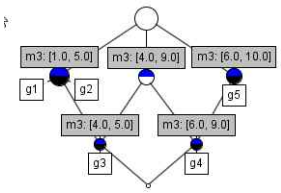
그림 4. 표8로부터 추출된 개념격자
Fig. 4. Concept Lattice extracted from Table 8



(a) θ = 3



(b) θ = 9



(c) θ = 5

그림 5. 각 속성들에 대한 개념격자
Fig. 5. Concept Lattices for each attributes

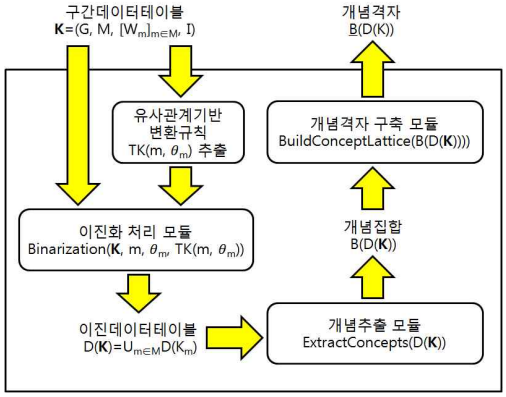


그림 6. 지원도구(iFCA)의 구성도
Fig. 6. Architecture of Supporting Tool(iFCA)

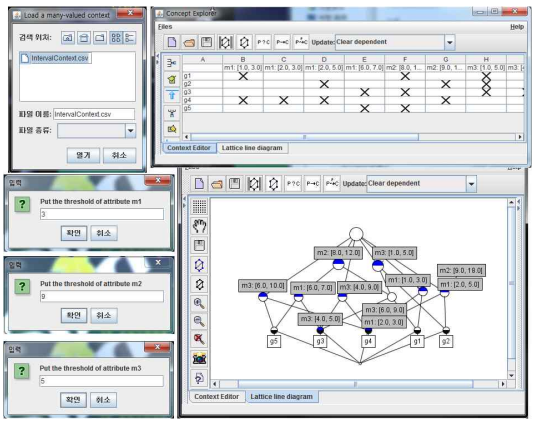


그림 7. iFCA의 실행화면
Fig. 7. Screen Shots of iFCA

그림5의 (a), (b), (c)는 속성m1, m2, m3 각각에 대한 객체g1~g5의 분류체계를 보여주고 있다. 예를들어, 그림 5(a)는, 주어진 임계값Θ=3과 속성m1에 대하여, iFCA를 적용하여용하여 객체들을 분류한 결과, 객체g1은 유사클래스 [1,3]에, 객체g3과 g5는 유사클래스[6,7]에 분류됨을 나타내고 있다. 또한, 그림5(b)에서는, 속성m2에 대하여 임계값 Θ=9에 의해, 각 객체들이 2개의 그룹(객체g2와 g4가 유사클래스[9.0, 18.0]에 속하고, g1, g3, g5는 유사클래스 [8.0, 12.0]에 속함)으로 분류되고 있다. 한편, 그림5(c)의 경우에는, 속성m3에 관하여 5개의 그룹으로 분류되고 있음을 알 수 있다.

	Price	Engine Capacity	Top Speed	Step	Length	Width	Height
car_1	[27806,33596]	[1370,1910]	[185,211]	[254,254]	[406,406]	[171,171]	[143,143]
car_2	[40230,68838]	[1595,1781]	[189,238]	[250,251]	[415,415]	[174,174]	[142,142]
car_3	[19229,30885]	[1242,1910]	[155,170]	[246,246]	[380,384]	[166,166]	[148,148]
car_4	[19242,24742]	[1242,1753]	[167,167]	[245,245]	[383,383]	[163,163]	[132,132]
car_5	[19837,29034]	[1242,1242]	[158,174]	[238,238]	[372,372]	[169,169]	[144,144]
car_6	[18492,24192]	[998,1348]	[150,164]	[236,236]	[375,375]	[160,160]	[144,144]
car_7	[19212,30612]	[973,1796]	[155,202]	[249,249]	[382,382]	[165,165]	[144,144]
car_8	[16992,23492]	[1149,1149]	[151,168]	[235,235]	[343,343]	[163,163]	[142,142]
car_9	[21492,33042]	[1119,1994]	[160,185]	[251,251]	[399,399]	[169,169]	[142,142]
car_10	[19519,32686]	[1397,1896]	[157,183]	[246,246]	[396,396]	[165,165]	[145,145]
car_11	[41593,62291]	[1598,2492]	[200,227]	[260,260]	[443,443]	[175,175]	[142,142]
car_12	[68216,140265]	[1781,4172]	[216,250]	[276,276]	[480,480]	[181,181]	[145,145]
car_28	[240292,391692]	[3586,5474]	[295,298]	[260,260]	[476,476]	[192,192]	[130,130]
car_29	[205242,215242]	[2977,3179]	[260,270]	[253,253]	[414,414]	[175,175]	[129,129]
car_30	[413000,423000]	[5992,5992]	[335,335]	[265,265]	[447,447]	[204,204]	[111,111]
car_31	[155000,159500]	[3217,3217]	[280,290]	[266,266]	[451,451]	[182,182]	[131,131]
car_32	[132800,262500]	[2799,5987]	[232,250]	[252,252]	[447,447]	[181,181]	[129,129]
car_33	[147704,246412]	[3387,3600]	[280,305]	[235,235]	[443,444]	[177,183]	[130,131]

그림 8. 실험데이터
Fig. 8. Experiment Data

표 9. 그림8의 실험데이터에 대한 주요한 3가지 속성들의 임계값
Table 9. Threshold values for the major three attributes of experiment data in Fig. 8

Θm	Price	Engine Capacity	Top Speed
I	4500.0	0.0	0.0
II	168149.58	2389.43	68.24
III	10000.0	1000.0	10.0

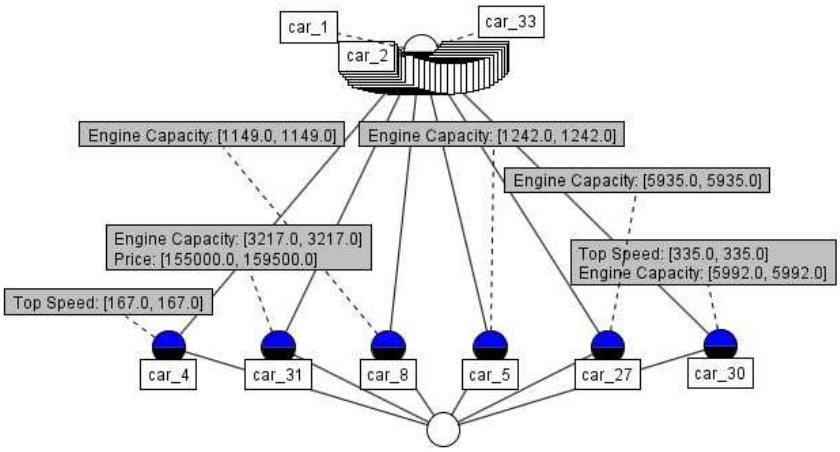


그림 9. 표9의 임계값 집합을 3가지 속성(Price, Engine Capacity, Top Speed)에 적용하여 생성된 분류체계
Fig. 9. Classification Hierarchy generated by applying the threshold value set I of Table 9 on the three attributes(Price, Engine Capacity, Top Speed)

Ⅳ. 지원도구 iFCA의 개발과 적용사례

그림6은 본 논문에서 제안한 구간데이터의 유사관계 $\simeq_\theta^m$ 를 기반으로 하는 형식개념분석기법을 지원하는 도구(iFCA)에 대한 구성도를 나타내고 있다. iFCA는 유사관계기반 변환규칙 추출모듈, 이진화처리모듈, 개념추출모듈, 개념격자구축 모듈 등, 총 4개의 모듈들로 구성되어 있으며, 그림7의 iFCA 실행화면과 같이 다음과 같은 처리흐름을 갖는다. 즉, 주어진 구간데이터는 구간데이터테이블 형태로 입력되어, 유사관계 $\simeq_\theta^m$ 를 기반으로 변환규칙 TK(m, Θ)이 추출된다. 변환규칙 TK(m, Θ)을 토대로 구간데이터테이블은 이진화처리모듈에 의해 이진데이터테이블로 변환된다. 이후, 개념추출모듈에 의해 개념들의 집합과 개념들 사이의 상하위관계가 추출되어, 마지막으로 개념격자구축 모듈에 의해 최종결과물로서 개념격자가 완성된다.

본 논문에서 제안한 구간데이터테이블에 대한 유사관계를 기반으로하는 형식개념분석에 의한 분류기법의 유용성을 확인하기 위하여 실제 데이터를 대상으로 지원도구(iFCA)를 적용한 실험결과를 보고한다. 이 실험에서는, 문헌[13,14]에서 언급되어 있는 구간데이터(http://lhedjazi.jimdo.com/useful-links에서 획득)중에서 자동차 관련데이터를 지원도구 iFCA의 입

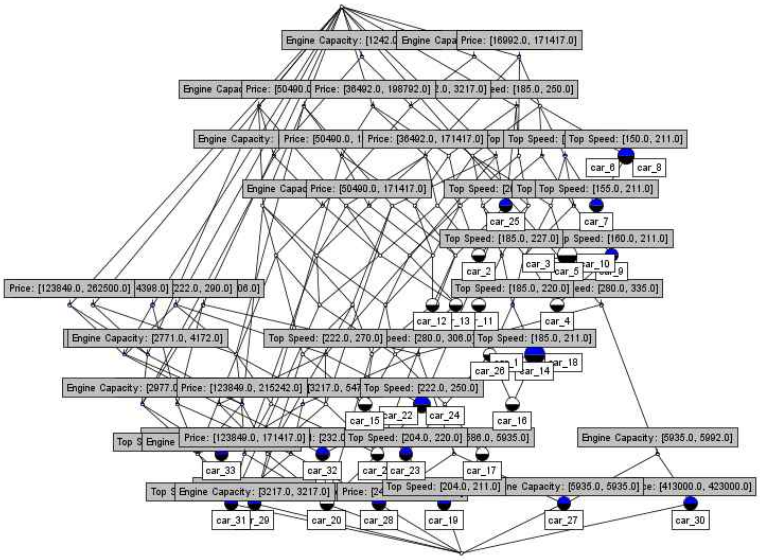


그림 10. 표9의 임계값 집합II를 3가지 속성(Price, Engine Capacity, Top Speed)에 적용하여 생성된 분류체계

Fig. 10. Classification Hierarchy generated by applying the threshold value set II of Table 9 on the three attributes(Price, Engine Capacity, Top Speed)

력으로 하였다. 이 실험에서 사용된 자동차관련데이터는 그림 8과 같이, 33대의 자동차에 대하여 가격, 엔진성능, 최고속력 등의 7개 속성들({Price, Engine Capacity, Top Speed, Step, Length, Width, Height})이 구간값으로 주어졌다.

각 구간데이터에 대한 실험에서는 각 속성m에 대한 임계값Θm을 어떻게 설정하느냐에 따라서 최종적으로 서로 다른 실험결과(즉, 분류체계를 나타내는 개념격자)가 생성된다. 본 실험에서는 실험데이터의 주요한 3가지 속성들(Price, Engine Capacity, Top Speed))에 대하여 구간속성값의 최소값과 평균값, 그리고 임의의 값을 각각 표9의 임계값Θm집합I, II, III으로 설정하여 분석을 수행하였다.

그림8의 자동차관련 구간데이터에 대하여 표9의 임계값집합 I, II, III을 토대로 iFCA를 이용하여 분석실험을 수행한 결과, 각각 58, 1182, 84개의 개념들이 추출되어 그림9~그림11과 같은 분류체계를 얻을 수 있다. 그림9의 분류체계로부터, Price, Engine Capacity, Top Speed속성의 임계값이 각각, 4500.0, 0.0, 0.0일 때, Top Speed속성값이 구간 [167.0, 167.0]에 속하는 자동차는 car_4이고, car_31은, Engine Capacity와 Price속성값이 각각 [3217.0, 3217.0], [155000.0, 159500.0]구간에 속하는 자동차로서 분류되었음을 알 수 있다. 또한, 표9의 임계값 집합II과 III을 적용한 실험결과는 그림10 및 그림11과 같다. 특히, 그림10의 분류체계에서 car_27과 car_30은 Top Speed가 [280.0,335.0]구간에 속하고 Engine Capacity가 [5935.0, 5992.0]구간에 속하는 것으로 분류되었으나, 그림11의 분류체계에서는, Engine Capacity(임계값=1000.0)의 최소값이 5935.0이고 최대값이 5992.0인 클래스에 속하는 자동차는 car_27과 car_30이고, car_27은 Top Speed의 구간값이 [298.0, 306.0]인 클래스에 속하며, car_30은 Top Speed가 [335.0, 335.0]이며 Price가 [413000.9, 423000.0]인 클래스에 속하는 것으로 분류된다는 사실을 알 수 있다.

위에서 수행한 실험결과로부터, 각 속성의 임계값을 구간값의 최소값에 근접시키면 대략적인 분류체계를 추출할 수 있고, 평균값에 근접시키면 보다 세밀한 분류체계를 얻을 수 있었다. 따라서, 분류목적 및 정밀도에 알맞도록 적정한 임계값을 설정하도록 해야하며, 어떠한 속성에 어떤 정도의 임계값을 설정하느냐에 따라서 주어진 구간데이터에 대하여 다양한 분류체계를 구축할 수 있음을 알 수 있다. 특히, 기존의 형식개념기법[7]에서는 구간데이터를 분석하기 위한 이진화처리과정이 제공되고 있지 않으므로, 위의 실험데이터를 토대로 적절한 분류체계를 도출할 수 없으나, 본 논문에서 제안한 분류기법을 적용하면 그림9~11과 같은 다양한 분류체계를 얻을 수 있다.

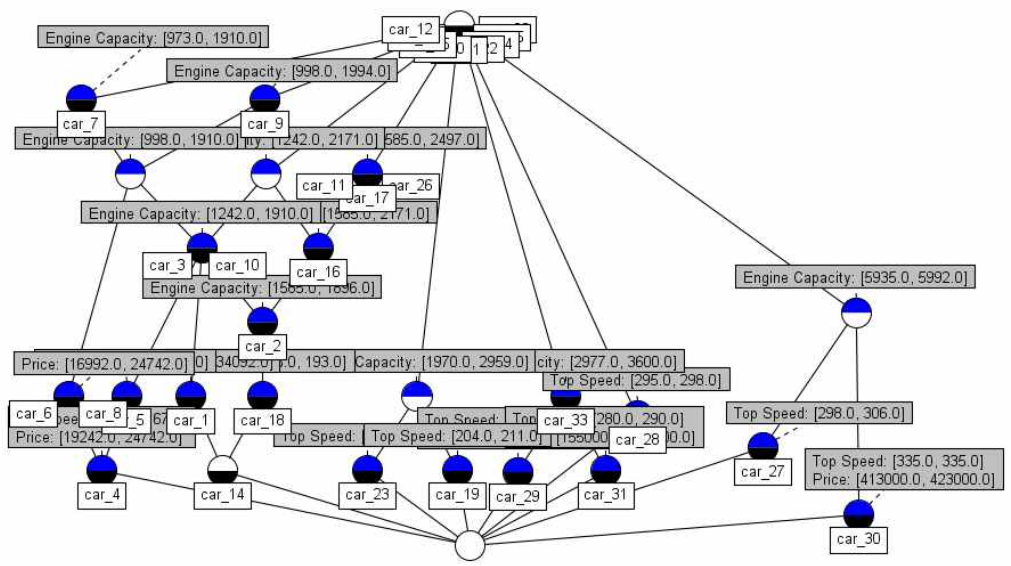


그림 11. 표9의 임계값 집합III을3가지 속성(Price, Engine Capacity, Top Speed)에 적용하여 생성된 분류체계
Fig. 11. Classification Hierarchy generated by applying the threshold value set III of Table 9 on the three attributes(Price, Engine Capacity, Top Speed)

V. 결론

본 연구에서는 다양한 유형의 데이터로부터 유용한 지식과 정보를 추출하기 위한 데이터마이닝기법으로서, 구간데이터의 유사관계 $\simeq_\theta^m$ 를 토대로 하는 형식개념분석기반의 분류기법을 제안하고, 이를 지원하는 도구(iFCA)를 개발하였다. 또한, 본 연구결과의 유용성을 확인하기 위하여 실제 구간데이터들에 대하여 지원도구iFCA를 적용함으로써, 구간값을 갖는 데이터에 대하여 다양한 형태의 군집화 및 분류체계를 구축할 수 있음을 확인할 수 있다. 특히, iFCA를 사용한 군집화 및 분류체계 구축에서는 임계값에 따라서 다양한 결과가 도출될 수 있으므로, 도구 사용자의 의도와 목적에 맞추어서 적절한 임계값을 설정해야만 효과적인 결과를 얻을 수 있다.

본 연구에서 제안한 데이터분석기법 및 지원도구는 일반적인 다치데이터의 분석에도 적용가능하며, 향후 여러가지 유형의 이진화기법들과 결합하여 다양한 유형의 데이터에 대한 군집화 및 분류체계화를 지원하는 종합적인 데이터분석기법으로 확장시켜서 데이터마이닝분야에서 유용하게 활용할 수 있을 것으로 기대한다.

참고문헌

[1] Pang-Ning Tan, Michael Steinbach, Vipin Kumar, "Introduction to Data Mining," Addison-Wesley, 2005.

[2] Ngai, E.W.T., Hu, Y., Wong, Y.H., Chen, Y. & Sun, X. "The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature." Decision Support Systems 50, pp.559-569, 2010.

[3] V. S. Verykios and E. Bertino and I. N. Fovino and L. P. Provenza and Y. Saygin and Y. Theodoridis, State-of-the-art in privacy preserving data mining, ACM SIGMOD Record, Vol. 1, No. 33, 2004.

[4] Clifton Phua and Vincent C. S. Lee and Kate Smith-Miles and Ross W. Gayler, "A Comprehensive Survey of Data Mining-based Fraud Detection Research," Artificial Intelligence Review, May 2005.

[5] Thabtah, Fadi Abdeljaber, "A review of associative

classification mining," Knowledge Engineering Review, Vol.22, No.1. pp.37-65, 2007.
[6] Ruotsalainen, Laura, Data Mining Tools for Technology and Competitive Intelligence, ESPOO 2008, VTT Tiedotteita ñ Research Notes 2451, 2008.
[7] Ganter, B., Wille, R., "Formal Concept Analysis: Mathematical foundations." Springer, 1999.
[8] Gerd Stumme, "Hierarchies of Conceptual Scales," Proceedings of Workshop on Knowledge Acquisition, Modeling and Management (KAW'99), 1999.
[9] J. Poelmans, P. Elzinga, S. Viaene, G. Dedene, "Formal Concept Analysis in Knowledge Discovery: A Survey," ICCS2010, pp.139-153, 2010.
[10] R. Cole, P. Eklund and D. Walker, "Using Conceptual Scaling In Formal Concept Analysis For Knowledge And Data Discovery In Medical Texts," International Symposium on Knowledge Retrieval, Use, and Storage for Efficiency, pp.151-164, 1997.
[11] Susanne Prediger, "Logical Scaling in Formal Concept Analysis," LNCS 1257, 1997.
[12] Joachim H. Correia, "Relational Scaling and Databases," Proceedings of the 10th International Conference on Conceptual Structures, LNCS2393, 2002.
[13] Lyamine Hedjazi and Joseph Aguilar-Martin and Marie-Véronique Le Lann, "Similarity-margin based feature selection for symbolic interval data," Pattern Recognition Letters, Vol.32, No.4, 2011.
[14] De Carvalho, F.A.T., De Souza, R.M.C.R., Chavent, M., Lechevallier, Y., "Adaptive Hausdorff distances and dynamic clustering of symbolic interval data," Pattern Recognition 27, pp.167 - .179, 2006.
[15] Quevedo, J., Puig, V., Cembrano, G., Blanch, J., Aguilar, J., Saporta, D., Benito, G., Hedo, M., Molina, A., "Validation and reconstruction of flow meter data in the Barcelona water distribution network." Journal of Control Eng. Practice 18, pp.640 - .651, 2010.

저 자 소 개

황 석 형

1991년8월: 강원대학교 전자계산학과 조기졸업(이학사).
1994년4월: 일본 오사카대학교 정보공학과 공학석사.
1997년4월: 일본 오사카대학교 정보공학과 공학박사
현 재: 선문대학교 컴퓨터공학과 교수
관심분야: 소프트웨어공학, 객체지향 분석/설계, Formal Concept Analysis, 온톨로지공학, 시맨틱 웹, Granular Computing 등
Email : shwang@sunmoon.ac.kr

김 응 희

2007: 선문대학교 컴퓨터정보학부 이학사.
2009: 서울대학교 치과대학 공학석사.
현 재: 서울대학교 의과대학 박사과정
관심분야: Data-mining, Formal Concept Analysis
Email : eungheekim@snu.ac.kr