개념계층과 속성감축에 기반한 효율적 데이터마이닝

김정호, 정 홍
phjhkim@hanmail.net
계명대학교 컴퓨터전자공학부

Efficient Data Mining Based on Concept hierachy and Attribute Reduction

Jung Ho Kim, Hong Chung
phjhkim@hanmail.net
Faculty of Computer & Electronic Engineering, Keimyung University

요 약

데이터베이스에서의 지식발견은 데이터베이스로부터 관심있는 지식을 발견하고 고수준의 언어로 지식을 표현하는 학습형태로서 여러 가지 기법들이 있으나, 단 하나의 기법의 적용으로는 각기 한계성 때문에 유용한 지식의 발견이 불충분하므로 이들의 특징을 잘 통합하고 발전시킨 새로운 기법이 필요하다.

본 논문에서는 데이터베이스의 일반화를 위한 개념계층의 상승방법과 불필요 속성의 감축 방법을 통합 적용함으로써 유용하고 간략한 최소 결정규칙을 자동적으로 생성하는 효율적 데이터 마이닝 방법을 제안한다.

1. 서론

데이터마이닝(data mining)은 데이터베이스나 정보 저장소에 있는 대량의 데이터에서 유용한 지식을 찾고자 하는 요구에 따라 많은 관심을 받고 있으며, 최근 대규모 데이터베이스에서 지식을 발견하기 위한 연구 및 개발 활동이 활발하게 이루어지고 있다[1].

지식발견 방법에 있어서 지식 감축(knowledge reduction)[2], 개념 계층(concept hierarchy)[3], 결정 트리(decision trees)에 의한 규칙 귀납[4] 등 상당한 연구가 진전되고 있으며, 또한 발견된 지식에 대한 추론방법의 개발이 진행되고 있다.

본 논문에서는 데이터베이스의 일반화를 위한 개념 상승(concept ascension)방법과 불필요 속성의 감축 방법을 통합 적용함으로써 유용하고 간략한 최소 결정규칙을 자동적으로 생성하는 효율적 데이터 마이닝 방법을 제안한다.

이를 위해 개념 상승에 의한 데이터베이스의 일반화, 속성의 중요도를 이용한 최적 감축, 속성값의 효율적인 감축 방법을 연구하고 이를 시스템으로 구현한다.

2. 개념계층과 개념상승

2.1 개념계층

대규모 데이터베이스는 일반적으로 대량의 튜플 뿐만 아니라 다수의 속성 및 속성값을 가지고 있다. 이와 같은 데이터베이스에서 각종 규칙 등 유용한 지식을 추출하기 위해 기초 속성값들을 일반화해야 한다[3]. 개념계층은 데이터베이스의 속성에 있어서 일반화 관계의 집합이다. 일반화 관계는 속성값의 전체집합과 이를 일반화한 단일값간의 관계이다. 즉, 속성 a의 일반화 관계는 a의 정의역이 $\{A_1,A_2,...A_k\}$이고 개념으로 표현된 단일값이 B일 때, $\{A_1,A_2,...A_k\}\subset B$로 표현되며, 이때 B는 각 $A_i(1\le i\le k)$의 일반화이다.

개념 계층은 자동적으로 또는 반자동적으로 구성할 수 있는데[3], 본 논문에서는 수치 속성에 대하여 완전 자동화가 가능한 클러스터링 방법을 사용하고, 비수치 속성에서는 실용적이고 간단한 전문가 지식을 이용하고자 한다. 수치 속성은 Fisher가 제안한 개념적 클러스터링 시스템인 COBWEB[5]에 의하여 자동적으로 조직화할 수 있는데, 이는 속성집합으로 기술된 객체를 분류 트리로 구성하는데 CU(Category Utility)라는 품질 척도를 사용한다. 즉, 클러스터 C를 n개의 상호배타적 클래스$C_1,..C_n$으로 분할하는데 있어서 CU는 분할 후 클래스 내의 유사성(intra-class similarity) 및 클래스간의 상이성(inter-class dissimilarity)을 의미하는 적합도(goodness)의 증가로 정의한다. 이 방법은 분류하는데 많은 메모리와 시간을 소요하므로 범주 데이터에만 적용되고 수치 데이터에는 적용하기 어렵다[6]. 본 논문에서는 Chu 등이 개발한 CoBase[6]에서 지식베이스를 구축하기 위한 TAH(Type Abstraction Hierarchies)의 생성에 CU를 근사적으로 계산하는 방법을 속성 단위 및 이진 분할 단위로 간략화하여 개념 트리의 자동생성에 사용한다. TAH에서는 클러스터링의 척도로서 RE(Relaxation Error)를 사용하는데, 클러스터 C가 x_i의 집합으로 구성되어 있을 때 실제 속성값과 일반화한 값간의 평균 차이로 정의한다.

$$\text{속성값 } x_i\text{의 } RE(x_i) = \sum_{j=1}^{n} P(x_j)|x_i - x_j| \qquad (2.1)$$

$P(x_j)$: C에서 속성값 x_j의 발생확률

$RE(x_i)$를 C의 모든 속성값 X_i에 대하여 합하면 다음과 같다.

$$\text{C전체의 } RE(C) = \sum_{i=1}^{n} P(x_i)RE(x_i) \qquad (2.2)$$

C의 분할 $P=\{C_1,..C_n\}$에서 분할 P의 RE는 다음과 같이 정의한다.

$$RE(P) = \sum_{k=1}^{n} P(C_k)RE(C_k) \qquad (2.3)$$

$P(C_k)$: C_k의 속성값 수를 C의 속성값 수로 나눈값

일반적으로 $RE(P)<RE(C)$인데, 이는 분할함으로써 RE가 감소함을 의미하므로, 최적 분할은 가장 적은 값을 갖는 RE(P)를 갖도록 분할한다. 그런데 하나의 클러스터를 n개의 서브클러스터로 분할하는 조합수는 n에 지수적이므로 최적분할 계산은 지수적 시간복잡도를 가진다. 따라서 본 논문에서는 계산 시간을 줄이기 위해 이진분할을 먼저하고, 이진분할 중 큰 서브클러스터를 또 이진분할하는 방법을 사용한다. 즉, 이진분할에서 시작하여 가장 큰 RE를 가지는 서브클러스터를 찾아 m개의 서브클러스터가 생성될 때까지 반복 이진분할 한다.

2.2 개념 상승

데이터베이스의 일반화인 개념 상승은 각 튜플의 속성값을 관련 속성의 개념 계층에서 상위수준의 개념으로 대치시킴으로써 수행된다[3]. 개념 계층의 상승은 데이터베이스가 일반화된 고수준의 개념을 가지며, 이때 중복되는 튜플은 합병하여 튜플 수를 줄인다.

개념이 상승된 일반화 데이터베이스에서 결정규칙을 도출할 때 조건속성은 동일한데 결정속성이 상이한 모순된 결정규칙이 생성되는 현상 즉, 결정속성에 대한 조건속성의 충돌이 발생할 수 있다. 이를 해결하기 위한 방법은 첫째 충돌이 발생한 튜플을 모두 제거하는 것인데, 이는 정보의 손실에 의하여 일부 규칙만 생성된다. 둘째 확률이 적은 튜플을 제거하는 것인데, 이는 규칙이 한쪽으로 편향되어 신뢰성이 결여된 규칙이 유도된다. 본 논문에서는 이를 모두 수용하는 방법을 사용한다. 즉, 모순된 두 개 이상의 규칙을 두 개 이상의 결정속성 값을 가지는 하나의 규칙으로 처리하여 각각의 결정속성값에 확률을

부여한다.

q를 일반화 튜플, Cj를 목적 클래스라고 하면, q에 대한 확률 Prob는 q에 의한 목적 클래스를 구성한 원 튜플의 수와 q와 동치인 모든 클래스에 있는 튜플 총수의 비율이다.

$$Prob = \frac{count(q \subset C_j)}{\sum_{i=1}^{K} count(q \subset C_i)} \quad (2.4)$$

count : 중복 튜플의 수, K : 클래스의 수,
C_j : $\{C_1,...,C_K\}$

개념 상승시 고려해야 할 또다른 문제는 빈도가 매우 적은 튜플의 처리인데, 이를 일반화 규칙으로 유도했을 때 규칙의 신뢰도가 매우 낮을 가능성이 있다. 따라서, 개념 상승관계에서 거의 나타나지 않는 튜플은 예외사항으로 간주하여 규칙의 일반화 이전에 사용자가 정한 잡음 필터 임계치보다 작을 때 제거한다.

q가 일반화 튜플일 때, q의 빈도를 Freq는 q의 중복 튜플 수와 총 튜플 수의 비율이다.

$$Freq = \frac{count(q)}{\sum_{i=1}^{n} count(q_i)} \quad (2.5)$$

잡음 필터 임계치는 일반화 관계에 있는 예외사항(매우 작은 빈도의 튜플)을 걸러내는 작은 값의 백분율이다.

3. 라프셋과 속성감축

3.1 라프셋

라프셋은 식별불가능(indiscernible) 객체의 클래스로 구성된 동치관계를 기본으로 한다[2].

U를 전체집합, R을 U에 있는 동치관계라 할 때, A=(U,R)을 근사공간이라 한다. x,y∈U, (x,y)∈R일 때 x와 y를 A에서 불분간이라고 한다. X를 U의 부분집합이라 할 때 A에서 X를 최소한 포함한 집합을 A에서 X의 상한근사라 하며 R_UX로 표기하고, A에서 X를 최대한 포함한 집합을 A에서 X의 하한근사라 하며 R_LX로 표기하며, 다음과 같이 정의한다.

A에서 X의 하한근사 : $R_LX = \{x \in U \mid [x]_R \subseteq X\}$

A에서 X의 상한근사 : $R_UX = \{x \in U \mid [x]_R \cap X \neq \emptyset\}$

여기서 $[x]_R$은 U의 원소 x에 대해 X를 포함하는 R의 동치클래스이다. 그리고 R_LX를 R에 대한 X의 양영역 $POS_R(X)$, $U-R_UX$를 음영역 $NEG_R(X)$, R_UX-R_LX를 경계영역이라 한다.

3.2 속성 감축

결정규칙 시스템에서 속성간의 관계를 분석하여 불필요한 속성을 발견하면 이를 제거함으로써 간략화할 수 있다. 즉, 간략화는 속성을 감축함으로써 이루어지는데, 속성 감축은 불필요한 속성을 제거하고 전체 속성집합과 같은 품질 척도를 갖는 최소의 부분 속성집합으로 정의한다[7].

S={U,A,V}에서 A=C∪D이고 B⊂C일 때 $POS_B(D)=POS_{B-\{p\}}(D)$이라면 D에 대해 속성 p∈B는 B에서 불필요(dispensable) 속성이고, 그렇지 않으면 필요(indispensable) 속성이다. 모든 p∈B가 필요 속성이면 B는 독립이다.

D에 대해 C에 있는 필수 속성집합이 C의 core로, 속성 감축에서 제거할 수 없는 속성이다.

$$CORE(C,D) = \{a \in C \mid POS_C(D) \neq POS_{C-\{a\}}(D)\}$$

B⊂C가 D에 대해 독립이고, $POS_C(D) = POS_B(D)$이면 B는 C에서 감축이다. D에 대한 C의 감축을 RED(C,D)라 할 때 $CORE(C,D) = \cap RED(C,D)$이다.

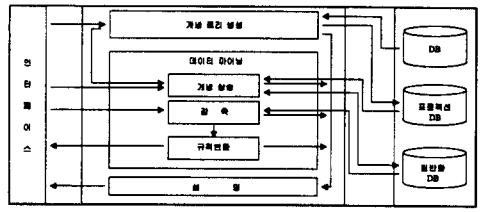
라프셋 이론에 의하면 모든 감축 가능집합에 대하여 감축 가능 여부를 조사해야 한다. 이는 지수적인 시간 복잡도를 가지므로 본 논문에서는 모든 경우의 조사가 아닌 휴리스틱(heuristic)한 방법으로 가장 좋은 하나의 감축을 찾아내는 방법을 사용한다. 이 방법은 Cercone[8]의 연구를 기반으로 한 것으로 가장 좋은 감축을 찾아내기 위해서 속성의 중요도 순서로 부분집합을 구성하여 가장 먼저 감축으로 형성되는 속성집합을 최적 감축으로 판단한다.

속성의 중요도를 계산하는 방법에는 통계학에서 사용하는 x^2 적합도 검증, 결정트리에 의한 기계 학습에서 사용하는 정보획득량 측정[9], 라프셋 이론에서의 속성간 중요도 계산 방법이 있는데, 이중 정보획득량 측정방법이 규칙의 발견에 있어서 우수하므로[10], 본 논문에서는 이 방법을 사용한다. 그리고 속성 감축시 core의 발견과 속성값의 감축은 식별가능 행렬(discernible matrix)과 식별가능 함수(discernible function)[11,12]를 사용한다.

4. 시스템의 구현

본 데이터 마이닝 시스템은 Windows98에서 VB5.0으로 구현하고 데이터베이스는 ACCESS97을 사용한다.

본 시스템은 그림 4.1과 같이 사용자 인터페이스, 개념 트리 생성 모듈, 데이터 마이닝 기관, 설명 모듈, 데이터베이스로 구성된다. 사용자 인터페이스는 각종 사용자 입출력과 실행과정을 대화식으로 처리하며 데이터 마이닝 기관은 개념 상승 모듈, 감축 모듈, 규칙변환 모듈로 구성된다.



<그림 4.1> 시스템 구조

· 개념 트리 생성 모듈 : DB로부터 관련 속성에 대한 프로젝션(projection) DB를 만들고 필요 속성별 개념 트리를 생성한다.

· 개념 상승 모듈 : 개념 트리를 사용하여 프로젝션 DB를 개념상승한 일반화 DB로 변환한다. 개념 상승수준 임계치와 예외사항을 제거하기 위한 잡음필터 임계치는 응용에 따라 사용자가 입력한다.

· 감축 모듈 : 일반화 DB로부터 불필요한 속성 및 속성값을 제거하여 최소규칙을 도출한다. 규칙유도를 위한 조건속성과 결정속성은 사용자가 입력한다.

· 규칙변환 모듈 : 최소규칙을 개념적인 언어로 표현한 일반화 규칙으로 변환한다.

· 설명 모듈 : 데이터 마이닝 과정에 있어서 사용자의 이해를 돕기 위해 필요한 부분에 대해 중간 실행과정을 출력한다.

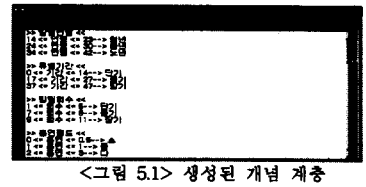
DB는 응용별 데이터베이스를 사용하며, 프로젝션 데이터베이스와 일반화 데이터베이스는 간단한 구조의 순차파일 형태로 구성하고 인터페이스는 대화식 메뉴 방식으로 구현한다.

5. 실험 및 평가

5.1 실험

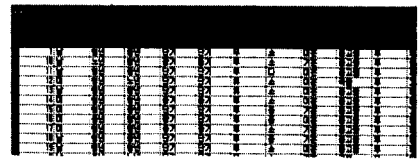
본 논문에서는 정신병의 진단에 따른 증상들을 중심으로 한 일반화 규칙을 유도해 보고자 한다. DB에 있는 속성은 성별, 나이, 발병연령, 유병기간, 입원회수, 가족력, 음주정도, 결혼유무, 학력, 자살시도, 종교, 흡연, 진단이다.

개념트리의 생성에 필요한 속성은 나이, 발병연령, 유병기간, 입원회수, 흡연이며, 자동으로 개념트리 생성이 가능하다. 개념트리 생성모듈의 수행후 생성된 개념계층은 그림 5.1과 같다.



<그림 5.1> 생성된 개념 계층

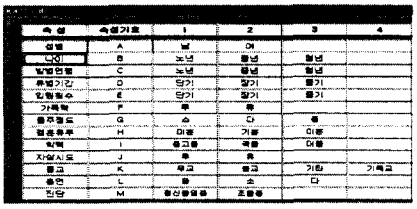
프로젝션된 DB에 개념 트리를 사용하여 개념 상승시켜 작성한 일반화 DB는 그림 5.2와 같다.



<그림 5.2> 생성된 일반화 DB

본 실험에서 진단을 결정속성으로 하고, 진단과 관련된 규칙

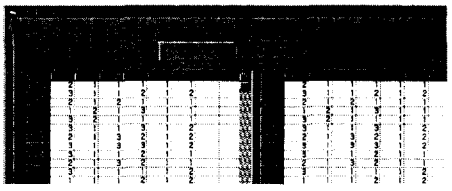
을 유도하기 위해 성별, 나이, 발병연령, 유병기간, 입원회수, 가족력, 음주정도, 결혼유무, 학력, 자살시도, 종교, 흡연을 조건속성으로 하였다. 편의상 속성 및 속성값은 그림 5.3과 같이 부호로 표현한다.



<그림 5.3> 부호표

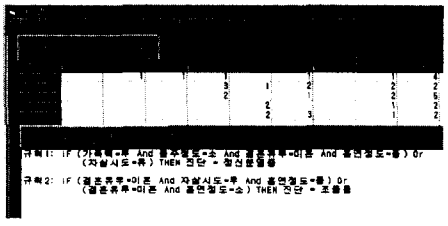
식별가능 행렬에서 단일 속성은 c, f, g, h, j, l이므로 core는 발병연령, 가족력, 음주정도, 결혼유무, 자살시도, 흡연이다. 그리고 core를 제외한 속성 a, b, d, e, i, k의 중요도를 계산하면 k(종교)가 0.992, d(유병기간)가 0.844, i(학력)가 0.821, e(입원회수)가 0.812, b(나이)가 0.777, a(성별)가 0.653이다.

core속성과 속성의 중요도를 사용하여 일반화 DB를 감축하면 그림 5.4과 같이 {c,f,g,h,j,l}이 최적 감축이다.



<그림 5.4> 최적감축

식별가능 행렬에서 객체별 식별가능 함수를 유도하고, 이를 흡수법칙에 의하여 간략화한 최소 규칙 집합은 그림 5.5과 같다.



<그림 5.5> 최소 규칙 집합

5.2 평가

본 논문에서 제안한 방법은 데이터베이스를 일반화시켜 추상성을 높일 뿐만 아니라, 속성의 중요도를 고려하여 감축을 생성하므로 감축 속도가 $O(n^2)$의 시간 복잡도를 가지며 또한 감축의 적합성을 판단할 수 있다.

그림 5.5에 있는 최소규칙을 테스트 데이터로 검증한 결과는 표 5.1와 같다. 테스트 데이터의 확률은 테스트 데이터를 각 규칙에 적용했을 때 진단에 일치하는 비율이다.

<표 5.1> 규칙의 검증

규 칙		진 단	테스트 데이터의 확률		
1	가족력=무 and 음주정도=소 and 결혼유무=미혼 and 흡연정도=중	정신분열증	정신	4	100%
	자살시도=유		정신	6	86 %
			조울	1	
2	결혼유무=이혼 and 자살시도=무 and 흡연정도=중	조울증	조울	1	100%
	결혼유무=미혼 and 흡연정도=소		정신	1	83 %
			조울	5	

표 5.1과 같이 본 논문에서 구현한 데이터 마이닝 시스템은 훈련데이터로부터 유도한 결정규칙이 테스트데이터에도 잘 적용됨을 보인다. 그런데 본 논문에서 사용한 데이터는 한 개 병원의 자료를 분석한 것이므로 유도된 규칙이 가장 일반화된 것이라고는 볼 수 없다.

6. 결론

본 논문에서는 특정 영역에 대한 지식을 일반화하고, 불필요한 사항을 제거하여 최소 결정규칙을 유도하였다. 이를 위해 클러스터링에 의한 개념트리 생성의 자동화, 개념 상승에 의한 데이터베이스의 일반화, 정보획득량 측정에 의한 속성의 중요도 계산, 중요도를 이용한 속성 감축에 의한 최적 감축, 식별가능 행렬을 이용한 효율적인 속성값 감축을 연구하고 프로토타이핑 시스템을 구현했다.

본 시스템은 첫째, 데이터베이스에 내재된 중요한 규칙을 발견하므로, 각종 투자 계획, 가격결정 등과 같은 의사결정 업무에 적용될 수 있다. 둘째, 데이터로부터 최적의 규칙을 유도하므로, 각종 고장진단, 의료진단 등의 전문가 시스템을 위한 지식베이스의 구축에 유효하게 사용될 수 있다. 셋째, 시장분석, 실험자료 분석 등 각종 데이터 분석에 이용될 수 있다. 그리고 데이터베이스의 정보검색에 있어서, 고수준 개념의 질의처리에 유용하게 이용 된다. 즉, 언어변수를 사용한 고수준의 질의를 개념 계층을 하향식으로 적용하여 적합한 튜플 집합을 검색할 수 있다.

데이터 마이닝은 최종 사용자의 지원이라는 목표를 추구하는데, 이는 의사결정 지원의 특성과 동질적이므로 본 논문의 결정규칙 유도 시스템을 의사결정 시스템과 통합된 환경으로 구축하도록 발전시키는 것이 바람직하다.

참고문헌

[1] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R. Uthurusamy (eds), *Advances in Knowledge Discovery and Data Mining*, AAAI/MIT Press, 1995

[2] Z. Pawlak, Rough Sets, *Theoretical Aspects of Reasoning about Data*, kluwer, 1991

[3] J. Han. Y, Cai, and N. Cercone, "Knowledge Discovery in Database: An Attribute-Oriented Approach," *Proceeding of the 18th Conference on Very Large Data Bases*, Vancouver, Canada, pp.340-355, 1992.

[4] J. Quinlan, "Induction of Decision Trees," *Machine Learning*, Vol. 1, No. 1, pp.81-106, 1986.

[5] D.Fisher, "Knowledge Acquisition via Incremental Conceptual Clustering," Machine Learning, vol. 2, no. 2, pp.139-172, 1987

[6] W. Chu, H. Yang, K. Chiang, M. Minock, G. Chow, and C. Larson, "CoBase: A Scalable and Extensible Cooperative Information System," *Intelligent Integration of Information*, G. Wiederhold ed. JIIS, Vol, 6, No. 2/3, pp.223-259, 1996.

[7] X. Hu, N. Cercone, and W. Ziarko, "Generation of Multiple Knowledge from Databases Based on Rough Set Theory," *Rough Sets and Data Mining*, T. Lin and N. Cercone eds, Kluwer, pp.109-121, 1997.

[8] N. Cercone, H. Hamilton, X, Hu and N. Shan, "Data Mining Using Attribute-Oriented Generalization and Information Reduction," *Rough Sets and Mining*, T. Lin and N. Cercone(eds), Kluwer, pp.199-277, 1997.

[9] M. Kamber, L. Winstone, W. Gong, S. Cheng and J. Han, "Generalization and Decision Tree Induction: Efficient Classfication in Data Mining," http://www.kdnuggets.com/, 1999

[10] 정홍, 최경욱, 정환묵, "Generation of Approximation Rules Using Information Gain," FUZZ-IEEE '99, The 8th Int'l Conf. on Fuzzy System, Seoul, Korea, Aug, '99. 22-25, 1999.

[11] A. Skowron and C. Rauszer, "The Discernibility Matrices and Fuctions in Information System," Slowinski (eds) *Intelligent Decision Support-Handbook of Advances and Applications of the Rough Set Theory*, Kluwer, pp.311-362, 1991.

[12] 이성주, 정환묵, 최규완, *러프집합과 응용*, 조선대학교 출판국, 1998.