통합 방법에 의한 지식 발견 시스템의 구현

김정호°, 정 홍
계명대학교 컴퓨터공학과

Implementation of Knowledge Discovery System Using Integrated Method

Jung-Ho Kim, Hong Chung
Department of Computer Engineering, Keimyung University

요 약

본 연구에서는 속성중심 귀납법에서 사용하는 개념 계층의 상승 기법, 결정트리에 의한 귀납법에서 사용하는 정보 획득량의 측정 기법, 그리고 라프셋에 의한 지식감축 방법을 복합하여 저수준의 데이터를 고수준 정보로 일반화하고, 불필요한 속성들을 감축하여 간략화된 결정규칙을 도출하는 통합방법의 지식 발견 시스템을 시험적으로 구현했다. 여기서 추출한 최소화 결정규칙은 대규모 데이터베이스에서 추출할수 있는 유용한 지식으로서 의사결정에 사용하는 정보가 된다 생성된 규칙지식은 각기 방법들보다 간결하다. 그리고 개념 일반화에 의해 유도된 지식이 고수준의 추상으로 표현된다

1. 서 론

데이터베이스에서의 지식 발견은 데이터베이스로부터 관심있는 지식을 발견하고 고수준 언어로 지식을 표현하는 학습 형태이다 KDD(Knowledge Discovery in Databases)에는 여러가지 기법들이 많이 있으나, 단 하나의 기법의 적용으로는 불충분하고 이들의 특징을 잘 결합한 통합 적용이 필요하다. 근래 많이 거론되고 있는 기법에는 Han[3]의 속성중심(Attribute-Oriented) 귀납법, Quinlan[7]의 결정트리(Decision Tree)에 의한 귀납법, Pawlak[5]의 라프셋(Rough Set)에 의한 지식감축 방법[6] 등이 많이 거론되고 있다 그런데 이들 기법들을 지식 발견에 단독으로 적용하기에는 문제점들이 있다. 속성중심 방법은 속성간의 종속관계를 분석하지 않아 생성된 규칙이 중복 정보와 불필요한 제약을 포함하고 있어 간략성이 없으며, 결정트리에 의한 방법과 라프셋 이론에 의한 방법은 속성과 튜플의 수가 많은 대형 데이터베이스에는 적용하기가 어렵다[4].

본 연구에서는 속성중심 귀납법에서 사용하는 개념 계층의 상승 기법, 결정트리에 의한 귀납법에서 사용하는 정보 획득량의 측정 기법, 그리고 라프셋에 의한 지식감축 방법을 복합하여 저수준의 데이터를 고수준 정보로 일반화하고, 불필요한 속성들을 최대한 감축하여 간략화된 결정규칙을 도출하는 통합방법의 지식 발견 시스템을 구현하고자 한다. 여기서 추출한 최소화 결정규칙은 대규모 데이터베이스에서 추출할수 있는 유용한 지식으로서 의사결정에 사용하는 정보가 된다

2. 데이터의 일반화에 의한 튜플 수의 감축

대규모 데이터베이스는 보통 속성값의 거대한 집합을 포함하고 있다 이를 간략화하기 위해서는 기초 데이터 인스턴스를 고수준으로 일반화해야 한다. 이 작업은 업무에 적합한 관계에 대한 속성중심 일반화에 의해 실현된다.

개념계층은 데이터베이스의 속성에 있어서 일반화 관계의 집합이다[8]. 일반화 관계는 속성값의 전체집합과 고수준으로 일반화된 단일값간의 관계이다 일반화 관계는 $\{a_1,a_2, \ldots ,a_K\} \subset A$로 표현되는데, A는 각 $a_i(1 \le i \le k)$의 일반화이다

일반화는 각 속성에 대한 정의역의 집합이 있고 고수준의 개념계층이 있으면 개념계층을 상승(각 튜플의 속성값에 대응하는 고수준 개념으로 대치)시킴으로써 수행된다 일반화시 키가 되는 속성은 제거한다. 이와같이 계념계층을 상승시켜 데이터베이스를 일반화하면 튜플 수를 줄일수 있다. 개념 상승에 의한 튜플 수의 감축에 대한 예는 [2]를 참고하면 된다

3. 정보 획득량에 의한 속성의 감축

결정속성에 영향을 적게 미치는 조건속성을 감축하기 위해 결정트리에 의한 귀납법중 정보 획득량의 측정방법을 이용한다

사례집합 K가 가지고 있는 정보값은 다음과 같은 엔트로피(Entropy)로 나타낼수 있다[1]

$$M(K) = \sum_{i=1}^{m} P_i \log_2(1/P_i) = - \sum_{i=1}^{m} P_i \log_2 P_i \qquad (1)$$

여기서 P_i는 클래스 K_i가 사례집합 K에서 차지하는 비율이다.

만약 정보값 M(K)를 가지고 있는 사례집합을 속성 X를 선택하여 하위 사례집합으로 나누었을 경우, 정보값 B(K,X)가 원래 M(K)보다 작다면 속성 X로 인해 정보값의 차이인 M(K)-B(K,X)만큼 정보를 획득한 셈이 된다

속성 X_j, j=1,...,m가 $|X_j|$가지의 속성값을 가지고, 클래스는 K_i, i=1,...,m 일 때, 속성 X_j를 사용하여 집합 K를 나누었을 경우 정보값 B(K,Xj)는 다음과 같다

$$B(K,X_j) = \sum_{i=1}^{|X_j|} Wi * M(Si) \qquad (2)$$

여기서 $M(S_i)$는 X_j 속성의 i번째 클래스의 값을 가지는 경우 하위 사례집합 S_i의 정보값이고, W_i는 가중치로서 다음과 같다

$$Wi = \frac{Si\text{에서의사례의수}}{K\text{에서의사례의수}}$$

상세한 예는 [2]를 참고하면 된다.

계산된 속성별 정보 획득량에 있어서 결정속성에 영향

을 적게 미치는 속성일수록 값이 적고, 영향을 많이 미칠수록 값이 크다 따라서 의미가 적다고 판단되는 속성을 제거해도 의사결정에 별로 영향을 미치지 않으므로 결정 테이블의 크기를 줄일수 있다.

4. 불필요한 속성값의 제거

불필요한 속성값을 제거하기 위해 라프셋 이론의 지식 감축 방법[6]을 이용한다.

U를 전체집합, R을 U에 있는 동치관계라 할 때, A=(U,R)를 근사공간이라 한다 x,y∈U, (x,y)∈R일 때 x와 y를 A에서 불분간이라고 한다 X를 U의 부분집합, $[x]_R$을 U의 원소 x에 대해 X를 포함하는 R의 동치 클래스라 할 때, 다음과 같이 하한근사(Lower Approximation)와 상한근사(Upper Approximation)를 정의한다

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

즉, $x_i \in R_LX$는 X에 확실히 포함되는 것이고, $x_i \in R_UX$는 X에 포함될 가능성이 있는 것이다

라프셋을 영역으로 정의하면 다음과 같다.

X의 R-양영역, $POS_R(X) = R_LX$
X의 R-음영역, $NEG_R(X) = U - R_UX$
X의 R-경계영역, $BND_R(X) = R_UX - R_LX$

다음과 같은 지식 시스템이 있을 때

$S=\{U,A,V\}$ $U=\{x_1,x_2,...,x_n\}$인 사례의 유한집합
$A=\{C,D\}$ 속성의 유한집합, C는 조건속성, D는 결정속성
$V=\bigcup_{p \in A} V_p$, V_p: 속성 p의 정의역

$POS_B(D)=POS_{B-\{p\}}(D)$라면 D에 대해 속성 $p \in B$는 B에서 불필요(dispensable) 속성이며, 그렇지 않으면 필수(indispensable) 속성이다

지식의 감축은 특정 지식에 필요한 기본 개념을 정의하는데 충분한 필수 부분이며, 코아(core)는 그중 가장 중요한 부분이다 D에 대해 C에 있는 필수 속성집합을 C의 코아라 하며 다음과 같이 정의한다.

$$CORE(C,D) = \{a \in C \mid POS_C(D) \neq POS_{C-\{a\}}(D)\} \quad (3)$$

$B \subset C$일 때 $POS_C(D)=POS_B(D)$이면 B는 지식 시스템의 감축이다 즉, 감축은 지식 시스템에 의해 결정규칙을 분간할 수 있는 필수 부분이다 그리고 D에 대한 C의 감축 RED(C,D)와의 관계는 다음과 같다

$$\cap RED(C,D) = CORE(C,D) \quad (4)$$

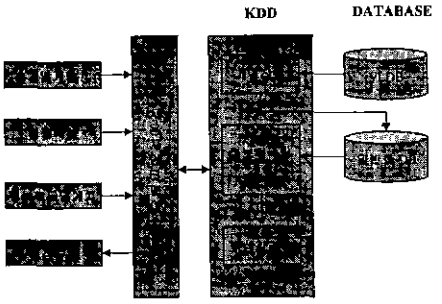
집합군 $F=\{X1, ,Xn\}$에서 $\cap\{F-\{Xi\}\}=\cap F$이면 X_1는 dispensable고, 그렇지 않으면 indispensable이다. 감축에 대한 상세한 예는 [6]을 참고하면 된다

5. 지식 발견 시스템의 구현

개념을 일반화하여 데이터베이스의 튜플 수를 줄이는 개념계층의 상승 방법, 결정트리의 정보 획득량 계산에 의하여 속성의 수를 줄이는 방법, 그리고 라프셋에 의한 속성값의 감축 방법을 통합 적용한 지식 발견 시스템을 VB50으로 구현하고 본대학 학적데이터베이스에 테스트하였다

5.1 시스템의 구조

본 시스템은 그림-1과 같이 사용자 인터페이스, KDD 처리기, 데이터베이스로 구성된다. 사용자 인터페이스는 각종 사용자 입력을 대화식으로 처리하며, KDD 처리기는 다음과 같은 3개의 모듈로 구성된다



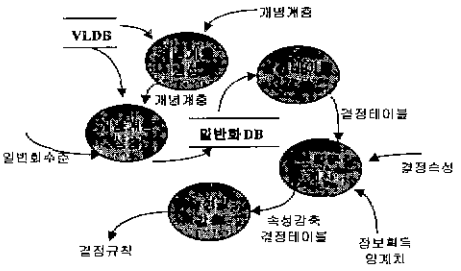
<그림-1>

- 개념 상승 모듈. 사용자가 작성한 개념계층을 대규모 데이터베이스에 적용하여 일반화된 소규모 데이터베이스를 생성한다.
- 정보 획득량 계산 모듈. 일반화된 소규모 데이터베이스를 결정 테이블로 변환시켜 속성별 정보 획득량을 계산하고 의미가 적은 속성은 삭제한다
- 속성값 감축 모듈 불필요 속성이 제거된 결정 테이블에서 속성값간 관계를 분석하여 불필요한 속성값을 제거하고 최소화한 결정 규칙을 생성한다.

데이터베이스는 응용별 데이터베이스를 사용하며, 일반화 데이터베이스는 간단한 구조의 순차파일 형태로 만든다

5.2 통합 시스템의 흐름도

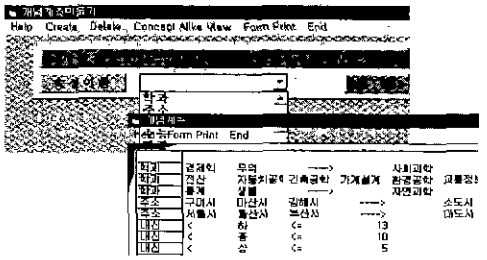
본 시스템의 흐름도는 그림-2와 같다.



<그림-2>

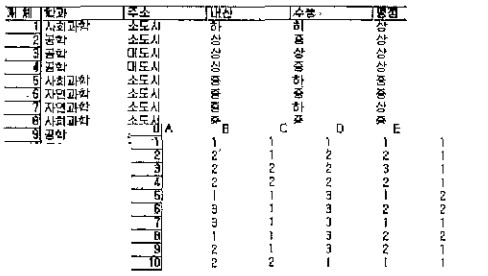
- 개념계층 작성. 사용자가 데이터베이스에 있는 속성을 분석하여 개념계층을 직접 입력하고 개념계층 트리를 구성한다 일반적으로 영역 전문가가 이 작업을 수행한다 개념계층의 작성 과정 및 작성된 개념계층의 보기는 그림-3과 같다.

- 개념계층 상승. 데이터베이스에 개념계층을 적용하여 일반화 데이터베이스를 생성한다 이때 개념을 어느 정도 상승시킬지는 사용자가 그 수준을 정한다. 일반화 데이터베이스에서 중복 튜플은 모두 제거한다.



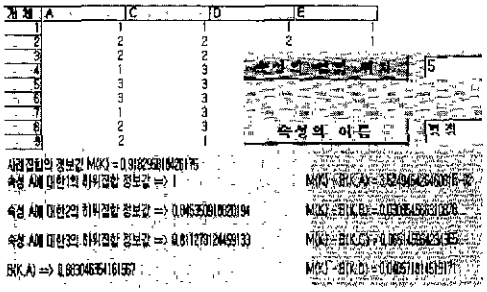
<그림-3>

- 결정 테이블 작성: 일반화 데이터베이스의 속성값을 수치화하여 데이터베이스를 결정 테이블 형태로 변형한다. 이는 이후 작업을 효율적으로 처리하고자 함이다. 결정 테이블의 보기는 그림-4와 같다.



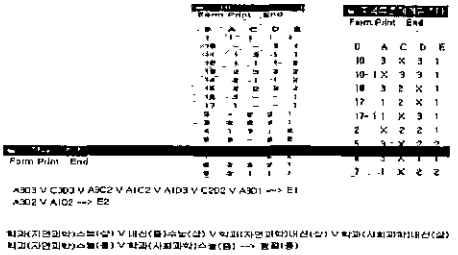
<그림-4>

- 정보 획득량 계산: 결정 테이블에서 결정속성을 지정하고 식 (1)과 (2)에 의하여 각 속성별 정보 획득량을 계산한다. 사용자는 정보 획득량을 보고 결정속성에 별로 영향을 미치지 않는 범위의 정보획득 임계치를 입력하면 임계치보다 적은 값을 가지는 속성은 제거되어 감축된 결정 테이블이 생성된다. 이 테이블에서 중복 사례가 있으면 제거한다. 감축 결정속성 지정, 속성별 정보 획득량 계산과정 및 결과는 그림-5와 같다.



<그림-5>

- 속성값 감축: 결정 테이블에서 속성간 속성값을 비교하여 식 (3)에 의해 사례별 코아를 계산하고 코아 테이블을 생성한다. 그리고 코아 테이블에서 식 (4)를 적용하여 사례별 감축을 계산하고 감축 테이블을 작성한다. 마지막으로 감축 테이블에서 최소화된 결정규칙을 도출한다. 코아 테이블, 감축 테이블 및 추출된 결정규칙의 예는 그림-6과 같다.



<그림-6>

6. 결 론

본 논문에서는 대규모 데이터베이스로부터 지식을 발견하기 위해 먼저 데이터베이스를 개념적으로 일반화하여 데이터의 크기를 줄이고, 결정속성에 영향을 적게 미치는 조건속성을 식제하여 속성의 수를 줄이며, 속성간의 종속관계를 분석하여 불필요한 속성값을 제거하는 방법으로 간략화된 형태의 지식을 도출한다. 이 방법은 속성중심 귀납법, 결정트리에 의한 귀납법, 라프셋에 의한 지식감축 방법 등 여러 가지 지식발견 방법중 중심이 되는 특성들을 결합하여 적용한 통합 방법이다.

이 통합 방법은 속성중심 방법과 결정트리에 의한 귀납법의 단점을 라프셋 이론의 지식감축 방법에 의하 해결했으며, 결정트리에 의한 방법과 라프셋에 의한 방법의 단점을 속성중심의 귀납법에 의하여 해결했다. 그리고 속성의 감축에 있어서 결정트리의 방법을 유효하게 사용했다 따라서 생성된 규칙지식은 각기 방법들보다 간결하다 그리고 개념 일반화에 의해 유도된 지식이 고수준의 추상으로 표현된다.

참고 문헌

[1] 이재규 등, 전문가 시스템, 법영사, 1995

[2] H. Chung, K.O. Choi, and H.M. Chung, "Integrated Method for Knowledge Discovery in Databases", The 3rd Asian Fuzzy Systems Symposium, June 18-21, Masan, Korea, pp 122-127, 1998

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

[4] X Hu, N Cercone, and J Han, "An Attribute-Oriented Rough Set Approach for Knowledge in Databases," Proceedings of the International Workshop on Rough Sets and Knowledge Discovery, Alberta, Canada, 12-15 October, pp 90-99, 1993

[5] Z. Pawlak, "Rough Sets", International Journal of Computer and Information Sciences, No.11, pp 341-356, 1982

[6] Z Pawlak, Rough Sets, Theoretical Aspects of Reasoning about Data, Kluwer, 1991

[7] J. R. Quinlan, "Induction of Decision Trees," Machine Learning 1, pp 81-106,1986

[8] D Fudger and H. Hamilton, "A Heuristic for Evaluating Databases for Knowledge Discovery with DBLEARN", Proceedings of the International Workshop on Rough Sets and Knowledge Discovery, Alberta, Canada, 12-15 October, pp.44-51, 1993