한국산학기술학회논문지
Vol. 11, No. 6, pp. 2214-2218, 2010

퍼지 연관규칙을 이용한 지능적 질의해석

김미혜[1*]
[1]충북대학교 전자정보대학

Intelligent Query Analysis using Fuzzy Association Rule

Mi-Hye Kim[1*]
[1]**College of Electrical & Computer Engineering, Chungbuk National University**

요 약 대용량 데이터에서 의미있고 유용한 지식을 추출하는 기법 중의 하나인 연관규칙은 데이터베이스에 존재하는 속성들 사이에 유사성 또는 패턴을 기술하여 사용자에게 데이터에 관한 유용한 정보를 줄 수 있다. 기존에 연구되어 온 연관규칙은 이진(boolean) 데이터베이스에 존재하는 유무에 대한 규칙으로 발견하는 것에 대해서 주로 연구되어왔다. 본 논문에서는 정량적 속성의 데이터를 기호적 속성 값으로 바꾼 후 연관규칙을 추출함으로써, 퍼지개념을 사용한 퍼지 연관규칙을 이용한 지능적 질의 처리 시스템을 제안하고자 한다.

Abstract Association rule is one of meaning and useful extraction methods from large amounts of data, and furnish useful information to user for data describing a pattern or similarity among attributes in database. Association rule have been studied about existence and nonexistence rule in boolean database. In this paper, we propose an intelligent query system using fuzzy association rule by extraction association rule changing a quantitative attribute data to a nominal attribute value.

Key Words : Association rule, Intelligent query , Fuzzy integral

1. 서론

데이터 마이닝(data mining)이란 대용량 데이터에서 의미있고 유용한 지식을 추출하는 과정을 의미한다. 대용량 데이터베이스 분야에 기존의 인공지능, 기계학습의 연구를 접목시키는 분야로 다양한 지식패턴을 데이터베이스에서 추출하고 응용하는 지식탐사(knowledge discovery and data mining)가 생겨났다. 데이터 마이닝은 사용되는 지식 추출 방법에 따라 얻을 수 있는 지식의 형태가 다양하다. 대표적인 것으로는 클러스터링(clustering) 정보나 클래스의 특성추출(class characterization), 연관 규칙(association rule) 추출등을 들 수 있다.

상품 혹은 서비스와 같은 데이터 간의 관계를 살펴보고 이로부터 유용한 규칙을 찾아내고자 할 때 이용될 수 있는 기법 중의 하나인 연관규칙은 데이터베이스에 존재하는 속성들 사이에 유사성 또는 패턴을 기술하여 사용자에게 데이터에 관한 유용한 정보를 줄 수 있다. 예를 들어 '넥타이를 구입하는 고객은 셔츠도 구입한다.' 또는' 정장과 벨트를 구입하면 코트도 함께 구입한다.' 라는 규칙(rule)을 발견한다면 전혀 관계가 없다고 생각되는 항목들 간의 숨어있던 관계를 발견한 것이다. 따라서 데이터 안에 존재하는 항목간의 종속관계를 찾아내는 작업으로 제품이나 서비스의 교차판매(cross selling), 매장진열(display), 첨부우편(attached mailings)등의 다양한 분야에서 활용되어 왔다. 연관규칙은 처음 소개된 후 대용량의 데이터베이스에 응용되기 위한 기법을 적용하는 등의 연구방향, 성능위주로 고려한 해싱이나, 통계적인 정보를 이용하는 연구방향, 그리고 연관규칙의 표현력을 높여보려는 연구 등으로 진행되어 왔다[1-4]. 또한 연관규칙의 표현력을 높이려는 연구 등이 있다[5]. 그러나 기존에 연구되어 온 연관규칙은 소비자가 어떠한 물품을 샀는지에 대한 일종의 이진(boolean) 데이터베이스에 존재하는 구

"이 논문은 2008년도 충북대학교 학술연구지원사업의 연구비지원에 의하여 연구되었음"
*교신저자 : 김미혜(mhkim@cbnu.ac.kr)
접수일 10년 05월 20일 수정일 10년 06월 08일 게재확정일 10년 06월 18일

입여부만을 규칙으로 발견하는 것에 대해서 주로 연구되어왔다. 그에 못지않게 어떤 물품을 얼마만큼 구입하였는지에 대한 정량적(수치적, quantitative)에 대한 데이터에서도 소비자 성향을 분석하는데 좋은 규칙을 얻을 수 있는 연구가 미비하였다. 이유는 정량적인 데이터에서 연관규칙을 추출하기 위해서는 정량적인 데이터를 기호적(nomal) 속성 값으로 바꿔주는 일이 필요하지만 쉽지 않은 일이다. 예를 들어 "나이"에 대한 속성 값이 "30세"라면 "30세"를 "나이가 적다"의 속성 값으로 변환 할 것인지 "나이가 많다"의 속성 값으로 변환 할 것인지에 따라 정량데이터를 기호적 속성 값으로의 변환이 오류없이 잘 변환되었는지가 틀려지기 때문이다. 또한 규칙 추출 후 어떤 규칙으로 결론을 도출하느냐에 따라 규칙적용이 잘 되었는지가 그렇지 않은지가 틀려진다. 연관규칙은 빈발(Frequent) 항목의 조합에 의해 규칙이 생성되기 때문에 세 개의 항목 조합으로 생성된 규칙은 두 개의 항목 조합으로 생성된 규칙을 포함하게 된다.

본 논문에서는 정량적 속성의 데이터를 기호적 속성의 데이터를 기호적 속성 값으로 바꾼 후 연관규칙을 추출함으로써, 퍼지개념을 사용한 퍼지 연관규칙을 이용한 지능적 질의 처리 시스템을 제안하고자 한다.

지능적 질의 처리란 사용자의질의에 대한 답을 데이터베이스에서만 추출하는 것이 아니고, 기존의 지식을 활용하여 부가적이거나 참고적인 답을 같이 제출하거나, 질의 자체를 수정하여 효과적인 의사결정을 돕는 질의처리 방식이다.

2. 기존연구

연관규칙은 '$A \rightarrow B$' 의 형태로 표현하는데, "항목 A를 구입한 고객은 항목 B를 구입한다"는 뜻으로 해석할 수 있다. $I = \{i_1, i_2, \cdots, i_m\}$는 항목(item)들의 전체 집합이다. 데이터베이스에 있는 항목 집합(T)은 $T \subseteq I$인 집합이다. 트랜잭션 T에는 TID 라는 유일한 식별자를 가지고 있다.

어떤 항목 집합을 X라 하고 있을 때, T가 X를 포함하면, 즉 $X \subseteq T$를 만족하면 '트랜잭션 T는 항목 집합 X를 포함한다' 또는 '지지한다'고 한다. 데이터베이스에서 항목집합 X를 지지하는 트랜잭션의 수를 항목집합 X에 대한 지지수(support count)라고 하고 즉, $X \subseteq T$에 대한 $|T|$을 지지수라 한다. 또한, 트랜잭션의 총 수 개수에 대한 X의 지지수의 비율을 X의 지지도(support)라고 한다. 미리 지정해준 최소 지지도를 만족하는 항목집합을 빈발 항목 집합(frequent or large item set)이라 한다. k 개의 항목으로 이루어진 빈발항목 집합을 빈발$k-$항목 집합(frequent k-item set)이라고 한다.

2-빈발항목 집합으로부터 연관규칙을 생성할 수 있는데 연관규칙이 생성되기 위해서는 사용자에 의해 미리 주어진 최소 신뢰도를 만족하여야 연관 규칙이 될 수 있다.

본 논문에서는 연관규칙의 형태는

"만약 X 가 A 이면 Y는 B이다"

와 같고, 연관규칙을 추출하기 위한 척도로는 지지도(support) 식(1)과 같다.

$$\text{지지도} < X, A > = \frac{\sum_{t_i \in T} \prod_{x_i \in X} \mathrm{Ma_j} \in \mathrm{A}(\mathrm{t_i}[\mathrm{x_j}])}{T} \quad (1)$$

단, T는 전체 레코드 집합을 의미하고, i는 자체 레코드 개수이며, X는 속성들의 집합, j는 속성의 개수를, A는 소속 함수들의 집합을 M은 소속의 정도를 의미한다. 식(1)에서 제시하는 지지도는 전체 데이터에서 속성에 대한 항목의 빈도수를 의미하고 사용자가 주게 되는 최소 지지도 이상의 값만이 규칙을 생성할 수 있는 빈발 항목이 될 수 있다. 연관규칙을 추출하기 위한 또 하나의 척도는 신뢰도(confidence) 식(2)이 사용된다.

$$\text{신뢰도} << X, A > < Y, B > > = \frac{\sum_{t_i \in T} \prod_{z_k \in Z} \{\mathrm{Mc_k} \in \mathrm{C}(\mathrm{t_i}[\mathrm{x_j}])\}}{\sum_{y_i \in T} \prod_{x_i \in X} \{\mathrm{Ma_j} \in \mathrm{A}(\mathrm{t_i}[\mathrm{x_j}])\}} \quad (2)$$

단, $Z = X \cup Y$ 이고, $C = A \cup B$,이며 k는 속성의 개수 이다. 식(2)에서 나타내는 신뢰도는 규칙의 조건절 항목을 만족하는 데이터 빈도에 대한 조건, 결론절 항목이 동시에 만족되는 빈도수를 의미한다.

정량 데이터를 이용하여 연관규칙을 추출하는 방법으로는 퍼지 연관규칙과 연과규칙이 있다. 퍼지 연관규칙과 연관규칙의 차이점은 정량데이터를 기호적 소속 값으로 변환 할 때 기호적 소속 값의 소속 정도(membership) 계산 방법과 추출된 여러 규칙에서 결론을 도출하기 위해 어떤 규칙을 적용하는지에 관한 추론(inference) 부분에서의 차이로 두 알고리즘을 설명할 수 있다.

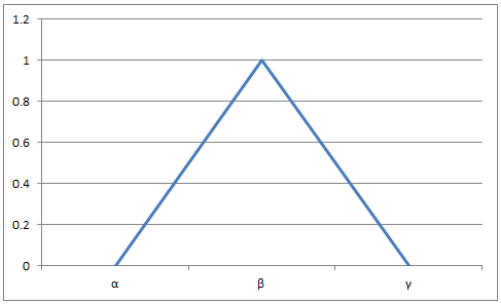
2.1 소속정도를 계산하는 방법

퍼지 연관규칙은 정량 데이터의 속성 값을 소속함수에 대한 소속정도로 변환될 때 소속정도가 0에서 1 사이의

값을 갖는다. 반면에 연관규칙은 소속정도가 0또는 1의 이진 값을 갖는다. 기존의 연구에서는 퍼지 연관규칙에서 소속정도를 계산하는 방법은 식(3), 식(4)과 같다.

$$(\alpha,\beta,\gamma)=\begin{cases}0 & \text{if} \quad x<\alpha \\ \dfrac{x-\alpha}{\gamma-\beta} & \text{if} \quad \alpha\le x\le\beta \\ \dfrac{\gamma-x}{\gamma-\beta} & \text{if} \quad \beta\le x\le\gamma \\ 0 & \text{if} \quad x>\gamma\end{cases} \tag{3}$$

삼각 소속함수 식(3)를 그림으로 표현하면 다음과 같다.

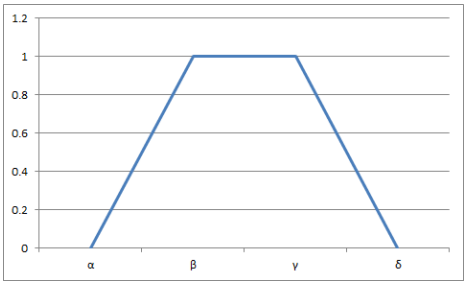


[그림 1] 삼각 소속함수

사다리꼴 함수 형태의 소속함수를 나타내는 수식은 식(4)와 같이 나타내며,

$$(\alpha,\beta,\gamma,\delta)=\begin{cases}0 & \text{if} \quad x<\alpha \\ \dfrac{x-\alpha}{\beta-\alpha} & \text{if} \quad \alpha\le x\le\beta \\ 1 & \text{if} \quad \beta\le x\le\gamma \\ \dfrac{x-\delta}{\gamma-\delta} & \text{if}\gamma\le x\le\delta \\ 0 & \text{if}x>\delta\end{cases} \tag{4}$$

이에 대한 소속함수 그림은 다음과 같다.



[그림 2] 사다리꼴 소속함수

2.2 추론하는 방법

추론은 입력데이터가 여러 규칙에 적용될 때 어떤 규칙을 이용하여 결론을 도출할 지에 관한 방법으로 퍼지 연관규칙은

$$\text{확신도}_{\,l}(t_i) = \max_{R(l)}(O_1\{\mathrm{Ma_j}\in\mathrm{A(t_i[x_j])}\}\cdot \mathrm{O_2}) \tag{5}$$

의해 적용하여 결론부를 적용할 규칙을 선택한다. 식(5)에서 R 은 규칙의 집합이며, l은 규칙개수를 의미하며 O_1 은 min 또는 product를 O_2= 지지도×신뢰도 또는 신뢰도를 의미한다. 반면 연관규칙은

$$\text{확신도}_{\,l}(t_i) = \max_{R(k)}\ O_2 \tag{6}$$

에 의해 적용하여 결론부를 적용할 규칙을 선택한다. 식(6)에서 R= 규칙의 집합을, l = 규칙개수를 O_2= 지지도X 신뢰도 또는 신뢰도를 의미한다.

주어진 데이터에 대해서 각 클래스에 대한 확신도(certainty factor)는 식(5)식(6)와 같이 계산되며 아래와 같은 표준 추론연산을 적용한다.

① min, product : 각 조건절의 소속정도를 합성하는 연산
② product : 각 규칙의 조건절의 소속정도를 합성한 결과와 그 규칙의 확신도(규칙의 지지도 X 신뢰도 또는 규칙의 신뢰도)를 합성하는 연산
③ max : 각 규칙의 결과를 합성하는 연산

3. 제안하는 알고리듬

정량적인 데이터를 통해 효율적인 연관규칙을 추출하기 위해서는 사용자의 개입이 필수적이다. 사용자의 개입이란 규칙추출을 위한 척도인 지지도, 신뢰도를 주는 것 이외에 연관규칙에서 가장 중요한 소속함수를 생성하는 것이다. 소속함수는 규칙의 성능과 이해성에 영향을 준다. 일반적으로 소속함수는 사용자(전문가)에 의해 주어지며 삼각형(triangular), 사다리꼴(trapezoidal), 가우시안(Gaussian) 함수 형태의 소속함수가 많이 사용된다. 본 논문에서는 퍼지적분을 이용하여 소속정도를 구하고자 한다. 우선 퍼지적분을 논의하기 전에 본 논문에서 사용될 몇 가지 정의와 기호를 먼저 소개하고자 한다.

Ω 는 공집합이 아닌 X의 부분집합의 σ -대수로 가정

하자. 그리고 다음과 같은 성질을 만족하는 집합치 함수 $\mu : \Omega \rightarrow [0, 1]$는 퍼지측도(fuzzy measure) 라고 불린다 ([10], [11]).

(1) $\mu(\varnothing) = 0$;

(2) $A, B \in \Omega$, $A \subset B$ 이면 $\mu(A) \leqq \mu(B)$;

(3) $A_n \in \Omega$, $A_1 \subset A_2 \subset \cdots$, 에 대해

$\lim_{n\to\infty} \mu(A_n) = \mu(\bigcup_{n=1}^{\infty} A_n)$ 성립하고,

(4) $A_n \in \Omega$, $A_1 \supset A_2 \supset \cdots$, 에 대해서

$\lim_{n\to\infty} \mu(A_n) = \mu(\bigcap_{n=1}^{\infty} A_n)$ 이 성립한다.

우리는 가측공간(measurable space) (X, Ω)에서 μ가 퍼지측도 일 때, (X, Ω, μ) 을 퍼지측도 공간(fuzzy measure space) 라고 부른다.

만약 B가 $[0, 1]$의 Borel 부분집합들의 σ -대수인 곳에서 $B \in$B 였을 경우, 어떤 B 값에 대해서도 $h^{-1}(B) = \{x \mid h(x) \in B\} \in \Omega$ 라면, 실수 값을 가지는 함수 $h: X \rightarrow [0, 1]$ 은 Ω와 B 에 대하여 Ω-가측(measurable) 이다. (단, 아무런 혼동이 없을 때 그냥 가측이라고 부르는 것이 가능하다.)

$L^0(X) = \{h: X \rightarrow [0, 1] \mid h$ 가 Ω와 B에 대하여 가측 (measurable) $\}$

와 같은 가측함수 집합을 생각 할 수 있다. 이 때, B는 통상적으로 $[0, 1]$의 Borel 부분집합들의 σ -대수이다. 주어진 h가 $h \in L^0(X)$인 어떤 경우에 있어서도, 우리는 $\alpha \in [0, 1]$일 때, $H_\alpha = \{x \mid h(x) \geq \alpha\}$라고 쓸 수 있다.

$A \in \Omega$ 이고, $h \in L^0(X)$라고 가정하자. μ에 대한 A 위에 있는 h의 퍼지적분(fuzzy integral)의 정의는 다음과 같다[10].

$$\int_A h \, d\mu = \sup_{\alpha \in [0, 1]} [\alpha \wedge \mu(A \cap H_\alpha)]$$

$A = X$일 때는 퍼지적분은 $\int h \, d\mu$ 로 나타내도 무방하다.

이러한 퍼지적분은 어떠한 대상을 종합적으로 평가할 경우 사용되는 경우가 일반적이다. 에에 의한 평가법의 알고리즘을 X가 유한 집합인 경우에 한하여 다음과 같이 제시 할 수 있다.

[단계 1] 유한집합 X의 원소로 정해진 평가항목에 대한 멱집합의 원소 $H \in P(X)$에 대해 평가기준의 중요도, 즉 ,전체집합 X에 대해 부분집합 H가 기여하는 정도 $\mu(H)$를 결정한다.

[단계 2] 각 평가 항목에 대한 평가값 $h(x_i)$을 $[0 , 1]$의 값으로서 구하고 그 크기 순으로 나열한다. 즉, $x_i \in X$ $(i = 1, 2, \cdots n)$에 대해

$h(x_1) \leq h(x_2) \leq \cdots \leq h(x_n)$ 이라 하고, 순서가 정해진 x_i들에 대해 H_i를 다음과 같이 구할 수 있다.

$H_i = \{x_k \mid k = i, i+1, \cdots, n\}$

[단계 3] 각각의 i에 대해서 $h(x_i) \wedge g(H_i)$를 계산한다.

[단계 4] 단계3에서 구한 모든 값의 max값을 소속정도로 한다.

데이터베이스의 질의는 사용자의 의도를 반영하여 전체 데이터베이스 중 관심 있는 부분을 지정하는 역할을 한다고 볼 수 있다. 질의내의 항목을 대상으로 연관 규칙을 응용하는 것이 가능하다면 사용자의 관심부분에 대한 규칙을 추출할 수 있다. 사용자 질의로부터 추출한 연관규칙은 지능적 질의처리에 활용된다. 기존의 지능적 질의처리로는 확장질의, 비교질의, 제안질의 등으로 사용자의 원래 질의에 부가적인 정보를 추가하여 사용자의 의사결정을 돕는다. 본 논문에서는 이러한 의사결정에 반영되는 사용자의 의도를 고려하여 퍼지적분을 이용한 소속함수로 소속정도를 나타내고자 하였다. 정량데이터에서 추출된 연관규칙의 인식율을 평가하는 기존의 연구방법으로는 사용자가 각 속성에 대하여 소속함수의 개수를 증가시킴으로써 인식율이 높이는 소속함수를 개수를 찾는 수동적인 방법을 사용하였다[6-8]. 가장 적절한 소속함수의 모양과 개수는 좋은 규칙을 생성하는데 가장 중요한 요소라고 볼 수 있다. 또한 정량 연관규칙에서 가장 중요한 요소인 소속함수의 모양을 히스토그램에 의해 자동 생성한다[8]. 히스토그램은 데이터 분포에 대한 통계적 특성을 나타내는 자료이므로 히스토그램을 이용하여 소속함수를 생성할 경우 효율적인 소속함수의 모양과 개수를 결정할 수 있다. 데이터의 각 속성에 대한 소속함수를 생성하기 위해서는 각 속성에 대해 클래스에 대한 히스토그램을 생성하여 히스토그램의 극대점과 극소점을 찾아 극대점에서 그 점의 양쪽 방향으로 가장 가까운 극소점을 직선으로 연결함으로 소속함수를 생성해서 기존의 연구들[6,7,8]과는 달리 퍼지적분을 활용한 소소함수 생성

은 보다 현실적인 알고리즘으로 간주된다.

4. 결론

지능적 질의 처리란 사용자의 질의에 대해 직접적인 응답만을 사용자에게 제공하는 것이 아니라 시스템이 보유하고 있는 정보를 활용해서 질의의 응답과 함께 보다 많은 정보를 제공 해 주는 것이다. 지능적 질의 처리를 위해 사용자 정보테이블이 존재해야한다. 사용자 정보 테이블에는 나이, 소속, 관심분야로 나누어진 그룹 정보가 있다. 사용자 정보 테이블은 이들 속성들에 정의된 개념 계층 트리를 이용한 미리 튜플 추상화 기법으로 처리된다. 일반화 트리로 추상화된 객체로 간주되면 20대 학생 -20대 전반 공과대학생- 23세 컴퓨터공학과 학부생 등과 같은 추상화 경로를 가지게 된다.

또, 속성들간의 계층 구조로는 학사관련 -학생 테이블- 학생의 나이와 같은 추상화 경로를 가지며, 이 계층은 시스템에서 제공하는 개념 계층 트리의 일종이다.

이러한 배경지식과 질의 패턴, 응답패턴에 대해 다시 튜플 추상화 기법을 적용하면, 유용한 연관규칙을 산출할 수 있다. 또, 이 때 필요한 패턴탐사 질의를 SQL과 유사한 형태로 정의한다. 이 경우 주어진 패턴 탐사 질의는 주변정보를 제공하는 지능질의 기법에 적용될 수 있다.

참고문헌

[1] R. Agrawal, S. Ghosh, T. Imielinski, B. Iyer and A. Swami, "An Interval Classifier for Database Mining Applications," Proc. 18th Int. Conf. VLDB pp.560-573, 1992.

[2] R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules", Proc. VLDB Cong. 1994.

[3] M. Houtsma and A. Swami, "Set-Oriented Mining for Association Rule in Relational Database", ICDE 95.

[4] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, and A. I. Verkamo, "Fast discovery of association rules," in U. M. Fayyad et. al. eds, Advances in Knowledge Discovery and Data Mining, Menlo Park, CA, AAAI/MIT Press, pp.307-328, 1996.

[5] Srikant, R. and R. Agrawal, "Mining Generalized Association Rules," Proc. 21th Int. Conf. VLDB, pp.407-419, 1995.

[6] 김준규, 안광일, 김성집, “수량과 시간 가중치를 고려한 퍼지 연관규칙 탐색방법,” 한국산업경영시스템학회, 춘계학술대회논문집 pp.395-401, 2004.

[7] 이동하, 김성민, 남도원, 이전영, “연관규칙을 이용한 지능적 질의처리 시스템,” 한국지능정보시스템학회, 한국지능정보시스템학회 학술대회논문집, pp.171-174. 1998.

[8] 손영경, 김명원, “퍼지 연관규칙과 연관규칙의 성능평가,” 한국정보과학회, 봄학술발표논문집, 제29권 제1호, pp.235-238, 2002.

[9] 장이채, "퍼지과학의 세계", 교우사, 1997

[10] M. Sugeno, "Theory of fuzzy integrals and its applications", Ph.D Dissertation Thesis, Tokyo Institute of Technology, 1974.

[11] Z. Wang, G. J. Klir, "Fuzzy Measure Theory," Plenum Press, New York, 1992.

김 미 혜(Mi-Hye Kim) **[정회원]**

- 1994년 2월 : 충북대학교 수학과 (이학석사)
- 2001년 2월 : 충북대학교 수학과 (이학박사)
- 2001년 4월 ~ 2004년 8월 : 충북대학교 초빙교수
- 2004년 9월 ~ 현재 : 충북대학교 전자정보대학 교수

<관심분야>
퍼지이론, 스테레오 비전, 제스츄어 인식