퍼지집합을 이용한 실수값 속성 사이에 존재하는 연관규칙의 발견

이지형○ 이광형
한국과학기술원 전산학과
305-701, 대전광역시 유성구 구성동
leejh@monami.kaist.ac.kr

Finding Association Rules among Real-valued Items using Fuzzy Sets

Jee-Hyong Lee Hyung Lee-Kwang
CS Dept. KAIST(Korea Advanced Institute of Science and Technology)
Kusong, Yousong, Taejon, 305-701, Seoul Korea
leejh@monami.kaist.ac.kr

요 약

연관규칙(Association Rule)은 데이타 베이스에 존재하는 속성들 사이의 관계를 기술하는 것으로, 간단하면서도 사용자에게 많은 정보를 줄 수 있다. 그러나, 지금까지는 이진 데이터 베이스에 존재하는 연관규칙의 발견에 대해서 주로 연구되어 왔으며, 실수값 속성을 갖는 데이타에 관한 연구는 미비하였다. 본 논문에서는 퍼지집합을 이용하여 실수값 사이에 존재하는 연관규칙을 기술하고, 그것을 찾아내는 방법을 제시한다. 제시하는 방법은 사용자에 의해서 정의된 언어항을 이용하여, 실수값 속성을 가진 데이타를 이진 데이타로 재구성한다. 그리고 재구성된 이진 데이타에 기존의 연관규칙 발견 방법을 이용하여 연관규칙을 찾아내고, 찾아진 연관규칙을 정의된 언어항을 이용하여 다시 기술한다.

1. 서 론

데이타 채광(data mining)은 쌓여있는 많은 데이타에서 현재까지 알려지지 않았고, 앞으로 유용할 것으로 생각되는 정보의 발견에대한 연구이다 [1] [2].

이러한 연구 중에서 연관규칙(Association Rule)은 간단하면서도, 사용자에게 중요한 정보를 준다. 연관규칙이란, 주어지는 데이타 투플에서 특정 속성의 나타남이 다른 속성의 나타남과 어떠한 관계를 갖고 있는가를 기술하는 것이다. 이러한 연관규칙은 간단한 형태로 다양하게 응용될 수 있다. 그러한 예로는 소비자들의 성향분석이 있다. 소비자 구매에 대한 데이타 투플을 분석하여 특정 품목을 구입하는 고객들이 어떤 다른 품목을 함께 구입하는 경우가 많음을 알아낸다면, 그러한 정보는 소비자 구매양식을 이해하는데 도움이 되며, 새로운 가치를 창출하는데 사용될 수 있을 것이다.

그러나 지금까지는 주로 이진 데이타 투플에 존재하는 연관규칙의 발견에 대하여 연구되었다. 그러나 일상에 존재하는 데이타들은 단순한 이진데이타가 아니라, 어느 정해진 실수영역의 값을 갖으므로, 이를 이진 데이타화할 경우에는 많은 정보의 손실이 생길 수 있다. 예를 들어, 소비자가 물건을 구입했을 때 생기는 데이타 투플은 단순히 특정물품을 구입했음만을 나타내지 않고, 소비자가 그 물품을 얼마나 많이 구입했나 하는 자료까지 포함하게 된다. 따라서, 주어지는 데이타에서 단순히 물건의 구입여부만을 이용한다면, 물건의 구입량은 고려하지 못하는 문제점이 있다.

이러한 점을 해결하기 위하여,본 연구에서는 퍼지집합을 이용하여 실수값을 갖는 데이타 사이에서 존재하는 연관규칙을 기술하는 방법과 그 발견 방법에 대하여 기술한다.

본 논문은 다음 같이 구성되어 있다. 2절에서는 기존의 연관규칙과 그 발견 알고리즘에 대해서 기술하고, 3절에서는 퍼지집합을 이용하여 확장된 연관규칙에 대해서 논의한다. 마지막으로 예제와 결론, 향후과제를 제시한다.

2. 연관규칙(Assocation Rule)

이 절에서는 이진데이타내에 존재하는 연관규칙과 연관규칙 발견 알고리즘에 대하여 기술한다 [2] [3]. 우선 속성(attribute) 집합 $R = \{I_1, I_2, \ldots, I_n\}$ 주어졌을 때, $t = \{t_1, \ldots, t_n\}, t_i \in \{0, 1\}$을 주어진 스키마 $\{I_1, I_2, \ldots, I_n\}$의 투플이라하고, 편의상 이것을 n개의 이진수로된 벡터로 생각한다. 그리고, t의 원

소중 속성 $I_i \in R$에 해당되는 값을 $t[I_i]$로 표현하면, $t[I_i] = 1$의 의미는 투플 t는 속성 I_i를 갖고 있음을 나타내게 된다. 그리고, $I \in W(W \subseteq R)$인 모든 I에 대하여 $t[I] = 1$인 경우, 이것을 간단하게 $t[W] = \overline{1}$로 표현한다. 또한 $W, V \subseteq R$; $A \in R$에 대하여 WV는 $W \cup V$를 WA는 $W \cup \{A\}$를 나타낸다. 그리고 $m(W)$는 $r[W] = \overline{1}$인 투플로 이루어진 집합으로, $|A|$는 집합 A의 원소의 갯수라고 정의한다.

연관규칙은 문법적으로는 $W \Rightarrow B, (W \subseteq R, B \in (R - W))$인 형태를 갖는다. 이것의 의미는 주어진 투플 중에서 $t[W] = \overline{1}$을 만족하는 것은 $t[B] = 1$도 만족한다는 것이다. 이것은 어느 투플에 속성집합 W가 나타날 때는 항상 속성 B도 나타났음을 의미한다. 그러나, 실제 데이터에서 $t[W] = \overline{1}$인 모든 투플이 $t[B] = 1$도 만족하는 경우는 매우 드물게 나타나므로, 일반적으로 다음과 같은 조건을 만족될 때 주어진 데이타 집합에서 연관규칙이 만족되는 것으로 한다.

n이 주어진 투플의 갯수일 때, 주어진 실수 γ, σ에 대하여

$$|m(WB)| \geq \sigma n \quad \text{이고,} \quad \frac{|m(WB)|}{|m(W)|} \geq \gamma$$

이면, 주어진 투플들은 연관규칙 $W \Rightarrow B$를 만족시킨다고 한다.

이때 γ를 확신임계값(condidence threshold), σ를 지지임계값(support threshold)이라 한다. 확신임계값은 한 연관규칙이 성립하기 위해서 필요한, 왼쪽의 속성과 오른쪽 속성 사이의 연관정도에 대한 최소값이고, 지지임계값은 한 연관규칙이 많은 수의 투플내에 존재해야 한다는 것의 기준이다. 즉, 데이타베이스 내에 그러한 속성을 가진 투플이 많은 경우에만 그것을 연관규칙으로 인정한다.

예를 들어 고객의 물품구입 데이타에서 다음 연관규칙이, $\gamma = 0.84, \sigma = 0.34$일 때 발견되었다면

햄버거 ⇒ 쥬스

이것은 고객의 34% 이상이 햄버거와 쥬스를 함께 샀으며, 햄버거를 산 고객 중 84%이상이 쥬스도 샀다는 것을 알려준다.

우리가 $|m(W)| \geq \sigma n$인 W를 포장집합(covering set)이라 한다면 연관규칙을 발견하는 기본적인 알고리즘은 다음 두 단계로 기술될 수 있다.

1. 주어진 지지임계값을 만족하는 모든 포장집합(covering set)을 찾는다.
2. 찾아진 포장집합에서 얻어질 수 있는 연관규칙을 조사하여, 확신임계값을 만족하는 것을 찾는다.

이것을 예를 들어 설명하면 다음과 같다. $R = \{A, B, C, D, E, F, G, H\}, \gamma = 0.9, \sigma = 0.3$이고, 다음과 같은 데이타 투플이 주어졌다면

	A	B	C	D	E	F	G	H
t_1	1	1	1	1	0	0	0	0
t_2	1	1	1	0	1	0	0	0
t_3	1	1	0	0	0	1	0	0
t_4	0	1	1	0	0	0	1	0
t_5	1	1	0	0	0	0	0	1

포장집합은 그 원소의 갯수가 1.5(=5 × 3)개 이상이어야 하므로, $\{A\}$, $\{B\}$, $\{C\}$, $\{A, B\}$, $\{A, C\}$, $\{B, C\}$, $\{A, B, C\}$이고, 이것들로부터 생성할 수 있는 연관규칙은 $\{A, B\}$에서 $A \Rightarrow B$, $B \Rightarrow A$, $\{A, C\}$에서 $A \Rightarrow C$, $C \Rightarrow A$, $\{B, C\}$에서 $B \Rightarrow C$, $C \Rightarrow B$, $\{A, B, C\}$에서 $AB \Rightarrow C$, $AC \Rightarrow B$, $BC \Rightarrow C$ 이다. 이 중에서 확신임계값을 넘는 것은 $A \Rightarrow B$, $C \Rightarrow B$, $AC \Rightarrow B$가 된다. 따라서, 주어진 데이타에서 $\gamma = 0.9, \sigma = 0.3$일 때 발견되는 연관규칙은 위의 세개가 된다.

그러나 이러한 연관규칙의 발견 방법은 속성의 갯수와 투플의 개수에 따라 크게 영향을 받고, 시간복잡도가 매우 커서, 규칙의 발견에 많은 시간이 걸리는 문제가 있다. 그러나 최근에는 이를 해결하고자 하는 많은 연구들이 진행되고 있으며, 좀 더 빠른 방법도 제시되고 있다 [5] [6].

3. 퍼지집합을 이용한 연관규칙의 확장

지금까지 기술한 연관규칙은 한 투플이 이진 데이터일때만 적용 가능하다. 이 절에서는 단순히 이진데이타에 적용되는 연관규칙이 아니라, 속성값이 실수일 경우에도 적용될 수 있는, 퍼지집합을 이용한 확장된 연관규칙에 대하여 제안한다. 기존의 연관규칙이 단순히 특정 속성이 어느 속성과 함께 나타나는 빈도가 높은가만을 기술하는 반면, 제안하는 확장은 어느 속성의 어느 값이 어느 속성의 어느 값이 함께 나타나는 빈도가 높은가를 기술하게 된다.

제안하는 방법은 주어지는 데이타내에 존재하는 연관규칙을 전문가나 사용자로부터 주어지는 언어항을 이용하여 기술한다. 주어진 데이타로부터 연관규칙을 발견하기 위하여, 제안하는 방법은 실수값 투플을 각 언어항의 소속도 투플로 바꾸고, 소속도 투플로부터 다시 이진데이타 투플을 얻어낸다. 구해진 이진투플에 기존 연관규칙 발견방법을 적용하여 연관규칙을 찾고, 이를 언어항을 이용하여 다시 기술하게 된다. 다음에서는 각 단계에 대하여 자세히 기술한다.

3.1 속성의 언어항

이진데이타로 이루어진 투플에서는 각 속성이 취할 수 있는 값은 $\{0, 1\}$이나, 속성이 제한된 범위 내의 실수값을 취하게 되는 경우에는 무수히 많은 수의 값을 갖을 수 있다. 따라서 특정 속성의 어느 값이 어떤 속성의 어떤 값과 함께 나타나는 빈도가 많은가를 확인하기 위해서, 주어지는 실수값을 그대로 사용하기는 매우 어렵다.

이러한 점을 고려하여 주어진 데이타를 직접사용하지 않고, 그 값을 사전에 정의된 몇개의 개념으로 일반화하여 사용하였다. 즉, 어느 속성이 취할 수 있는 값의 영역에 몇개의 개념을 나타내는 언어항과 퍼지집합을 정의하여, 그 속성값이 만족하는 언어항을 새로운 속성값으로 사용하고 각 속성의 언어항사이에 존재하는 연관규칙을 탐색하게 된다. 이렇게 그 속성값 자체를 이용하지 않고, 사용자나 전문가가 사전에 정의한 몇가지 언어항을 이용하므로써, 그 속성의 특징을 일반화하고 요약할 수 있는 특

징을 갖는다.

이를 위해서 본 논문에서는 각 속성마다 그 영역에 정의된 언어항과 그에 해당하는 퍼지집합이 있다고 가정하였다. 이것은 사전에 전문가 또는 사용자로부터 주어질 수 있으며, 이 때에 사용자는 자신의 관심과 속성의 중요도에 따라서 퍼지집합의 갯수와 위치를 결정할 수 있다.

이렇게 사용자는 자신이 갖고 있는 투플의 속성 중 실수값을 갖는 것들에 대하여 자신의 관심정도와 속성의 중요도에 따라서 퍼지집합의 위치와 갯수를 정하므로써, 사용자가 연관규칙의 발견에 개입할 수 있다는 장점을 갖는다. 즉, 사용자가 각 속성의 퍼지집합의 갯수와 위치를 변화시키면 이것은 후에 발견될 수 있는 연관규칙의 종류와 형태 및 규칙을 기술할 때의 자세함 정도를 변화시키게 된다.

3.2 소속도투플 생성

제안하는 방법은 각 속성값을 정의된 언어항을 이용하여, 주어진 투플을 재구성하게 된다. 이때 재구성된 투플은 각 속성의 값이 각 언어항을 어느 정도 만족하는가를 나타낸다. 이를 소속도(membership degree) 투플이라 정의한다.

소속도 투플은 다음과 같은 절차로 생성한다. 처음에 주어진 스키마 $R = \{I_1, I_2\}$가 있었고, 언어항이 I_1에는 $f^1_{I_1}, f^2_{I_1}, f^3_{I_1}$, I_2에는 $f^1_{I_2}, f^2_{I_2}$가 정의되었다면, 소속도 투플은 다음 스키마의 인스턴스가 된다.

$$M = \{\mu_{f^1_{I_1}}, \mu_{f^2_{I_1}}, \mu_{f^3_{I_1}}, \mu_{f^1_{I_2}}, \mu_{f^2_{I_2}}\}$$

즉, $t = \{t_1, t_2\}$이 주어져 있다면, 이것의 소속도 투플은

$$m = \{\mu_{f^1_{I_1}}(t_1), \mu_{f^2_{I_1}}(t_1), \mu_{f^3_{I_1}}(t_1), \mu_{f^1_{I_2}}(t_2), \mu_{f^2_{I_2}}(t_2)\}$$

가 된다. 이때, $\mu_f(x)$는 실수 x가 퍼지집합 f에 속하는 정도를 나타내는 소속도를 의미한다.

3.3 이진투플 생성

주어진 투플로부터 얻어진 소속도 투플에서 연관규칙을 찾아내기 위해서, 소속도 투플을 다시 이진투플로 바꾸게 된다. 이때 사전에 정의된 소속도 임계값(membership degree threshold) μ를 이용하게 된다. 주어진 소속도 투플의 모든값 중 μ보다 큰 값은 '1'로, 그보다 작은 경우는 '0'으로 바꾸어 준다. 예를 들면 $\mu = 0.3$이고, 소속도 투플이 $\{0.4, 0.6, 0, 0.8, 0.2\}$일때, 이진투플은 $\{1, 1, 0, 1, 0\}$이 된다. 각 속성의 언어항을 그 속성 영역에서 정의된 어떤 개념이라 했을때, 소속도가 μ 이상인 것에 대해서만 그 개념을 만족시키는 것으로 간주하게 된다. 이렇게 얻어진 이진 투플에 이미 언급된 연관규칙 발견기법을 적용하여 연관규칙을 찾게 된다.

3.4 확장된 연관규칙 생성

소속도 투플의 이진투플에 연관규칙 탐색기법을 적용하면, 소속도 투플의 각 속성사이에 존재하는 연관규칙을 찾을 수 있다. 이때 발견되는 연관규칙은 아래와 같이 두가지로 분류된다.

- 각 속성의 언어항 사이의 연관규칙
  (예:$\mu_{f^1_{I_1}} \Rightarrow \mu_{f^1_{I_2}}$)
- 한 속성의 언어항 사이의 연관규칙
  (예:$\mu_{f^1_{I_1}} \Rightarrow \mu_{f^2_{I_1}}$)

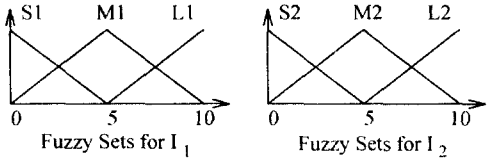
첫번째 연관 규칙은 어느 속성의 특정 값은 어느 속성의 어느 값과 함께 나타나는 빈도가 많음, 예를 들면 "속성 A가 큰값이면, 속성 B의 값도 크다" 등을 나타내지만, 두번째 연관규칙은 한 속성내의 연관, 즉 "속성 A가 큰값이면 속성 A가 중간값인 빈도가 높다"등을 기술하는 것으로 이것은 처음에 의도하였던 것이 아니며, 사용자에게 의미있는 정보도 주지 못한다. 두번째와 같은 연관규칙이 생성되는 이유는 이진 소속도 투플에 단순히 연관규칙 발견 방법을 그대로 적용했기 때문이다. 기존의 연관규칙 발견 방법은 모든 속성을 각각 독립된 속성으로 간주하나, 이진 소속도 투플은 몇개의 이진 속성이 하나의 실수 속성을 나타내기 때문에 이러한 결과가 생기게 된다. 따라서 이진 소속도 투플에서 발견될 수 있는 이러한 연관규칙을 배제해야 한다. 그리고 이러한 배제는 연관규칙의 발견을 좀 더 빠르게 할 수 있다.

그 다음 단계에서 찾아진 연관규칙을 언어항을 이용한 기술로 바꾼다. 예를 들면 찾아진 연관규칙에 $\mu_{f^1_{I_1}} \Rightarrow \mu_{f^1_{I_2}}$이 있다면, 이것은 어느 투플의 속성 I_1의 값이 퍼지집합 $f^1_{I_1}$을 만족시킬 때 속성 I_2은 퍼지집합 $f^1_{I_2}$을 만족시킨다는 것을 의미하므로, 이를 아래와 같이 바꾼다.

$$(I_1 \text{ is } f^1_{I_1}) \Rightarrow (I_2 \text{ is } f^1_{I_2})$$

지금까지 본 논문에서 제안하는 확장된 연관규칙과 그 발견방법에 대해서 설명하였다. 간단한 예를 살펴보면 다음과 같다.

$R = \{I_1, I_2\}$주어졌을때, 이것의 투플로 (8,7), (3,6)이 있고, I_1, I_2에 대한 언어항의 정의가 다음과 같이 동일하게 정의되어 있다면



이것의 소속도 스키마는

$$M = \{\mu_{S1}, \mu_{M1}, \mu_{L1}, \mu_{S2}, \mu_{M2}, \mu_{L2}\}$$

이다. 각 투플을 소속도 투플과 $\mu = 0.3$으로 했을때 이진소속도 투플은

$$(8,7) \rightarrow \{0, 0.4, 0.6, 0, 0.6, 0.4\} \rightarrow \{0, 1, 1, 0, 1, 1\}$$

$$(3,6) \rightarrow \{0.4, 0.6, 0, 0, 0.8, 0.2\} \rightarrow \{1, 1, 0, 0, 1, 0\}$$

이 된다. 이때, $\sigma = 0.5, \gamma = 0.9$로 하면 포장집합은 $\{\mu_{M1}\}, \{\mu_{M2}\}, \{\mu_{M1}, \mu_{M2}\}$이 되고, 연관규칙은 $\mu_{M1} \Rightarrow \mu_{M2}$, $\mu_{M2} \Rightarrow \mu_{M1}$이 된다. 이것을 언어항을 이용하여 기술하면, $(I_1 \text{ is } M1) \Rightarrow (I_2 \text{ is } M2)$과 $(I_2 \text{ is } M2) \Rightarrow (I_1 \text{ is } M1)$ 으로, 우리는 "속성 I_1이 중간정도이면, I_2도 중간정도를 갖다"는 것과 이것의 역을 연관규칙으로 얻을 수 있다.

4. 실험 및 결과

본 절에서는 제안하는 방법을 실제데이타에 적용하여 확장된 연관규칙을 탐색하고, 그 결과를 제시한다. 사용한 데이타*는 주택가에 관련된 것으로 모두 14개의 속성을 갖고 있으나, 실험에서는 이중 11개만을 사용하였다. 사용한 속성들은 모두 제한된 범위의 실수값을 갖는다. 사용된 데이타는 506개 지역의 학생/교사 비율(PTRATIO), 도매상의 사업영역의 비율(INDUS), 주요고용지역까지의 거리(DIS), 소유자가 살고있는 주택값 중의 중간값(MEDV), 25,000 sq.ft.보다 넓은 주거지역의 비율(ZN) 등에 대한 값으로 이루어져 있다. 실험에 사용한 소속도임계값은 $\mu = 0.2$, 지지임계값은 $\sigma = 0.2$, 확신임계값은 $\gamma = 0.95$이다. 각 속성마다 세개의 언어항(Zero, Medium, Large)을 정의하였다. 686개이며 다음은 발견된 연관규칙의 일부분이다.

- (INDUS is L) ⇒ (MEDV is L)
- (MEDV is L) ⇒ then (ZN is Z)
- (INDUS is M) (DIS is Z) (PTRATIO is L) ⇒ (TAX is L)
- (ZN is Z)(DIS is Z)(PTRATIO is L) (MEDV is L) ⇒ (INDUS is L)

첫번째와 두번째는 간단한 형태의 연관규칙이며, 세번째와 네번째는 좀 더 복잡한 형태의 연관규칙이다. 첫번째 연관규칙은 도매상의 사업영역이 넓은 지역의 주택값이 높다는 것을 의미하고, 네번째 규칙은 25,000 sq.ft.보다 작은 주거지역의 비가 낮고, 주요고용지역까지 거리가 가깝고, 학생/교사의 비가 높으며, 주택값이 높은 지역은 도매상의 사업영역이 넓다라는 것을 의미한다.

5. 결 론

본 논문에서는 이진데이타 투플에 존재하는 연관규칙의 기술및 발견 방법을 실수데이타 투플에 적용할 수 있도록 퍼지집합을 사용하여 확장하였다. 제안된 방법은 기존의 이진데이타에서 발견할 수 있는 모든 연관규칙을 포함하는 좀 더 일반화된 방법이다. 그리고 제안된 방법의 특징은 연관규칙의 발견과 기술에 사용자의 관심을 반영할 수 있다는 장점을 갖는다.

그러나, 대부분의 경우 많은 연관규칙이 발견되므로, 발견된 연관규칙을 그대로 사용하기에는 어려운 점이 있다. 따라서 발견된 연관규칙에서 중복성을 제거하고 의미있는 정보를 얻어낼 수있는 연구가 계속되어야겠다.

참고 문헌

[1] W.J. Frawley, G. Piatetsky-Shapiro, C.J. Matheus, *Knowledge Discovery in Databases: An Overview*, Knowledge Discovery in Databases(G. Piatetsky-Shapiro, W.J. Frawley, eds.), pp.1-27 AAAI Press/The MIT Press, US, 1993.

[2] 이도헌, *Data Mining Techniques : An Overview*, 한국정보과학회 추계학술발표회 특강요약집, pp.23-35, 1996.

[3] M. Holsheimer, M. Kersten, H. Mannila, H. Toivonen, *A Perspective on Databases and Data Mining*, the 1st International Conference on Knowledge Discovery and Data Mining, pp.150-155, 1995.

[4] M. Flemettinen, H. Mannila, P. Ronkainen, H. Toivonen, A. I. Verkamo, *Finding Interesting Rules from Large Sets of Discovered Association Rules*, the 3rd International Conference on Information and Knowledge Management, pp.401-407, 1994.

[5] H. Mannila, H. Toivonen, A.I. Verkamo, *Efficient Algorithms for Discovering Association Rules*, AAAI workshop on Knowledge Discovery in Databases, pp.181-192, Washington, July, 1994.

[6] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, A.I. Verkamo, *Fast Discovery of Association Rules*, Advances in Knowledge Discovery and Data Mining(U.M. Fayyad, et al. eds.), pp.307-328, AAAI Press/The MIT Press, US, 1996.

[7] E. Boros, P.L. Hammer, T. Ibaraki, A. Kogan, E. Mayoraz, I. Muchnik, *An Implementation of Logical Analysis of Data*, Rutcor Research Report(RRR #22-96), Rutgers Univ., July 1996.

*Merz, C.J., & Murphy, P.M. (1996). UCI Repository of machine learning databases [http://www.ics.uci.edu/~mlearn/MLRepository.html]. Irvine, CA: University of California, Department of Information and Computer Science.