퍼지 연관규칙과 연관규칙의 성능 평가

손영경[O] 김명원
숭실대학교 컴퓨터학과
saint75@hanmail.net, mkim@computing.soongsil.ac.kr

Performance Estimation of Fuzzy Quantitative Association Rules and Crisp Quantitative Association Rules

Young-Kyong Son[O] Myung-Won Kim
School of Computing, Soongsil Univ.

요 약

연관규칙(association rule)이란 데이터 베이스에 존재하는 속성들 사이에 유사성 또는 패턴을 기술하는 것으로, 사용자에게 데이터에 관한 유용한 정보를 줄 수 있다. 그러나, 지금까지의 연관규칙은 이진(boolean) 데이터 베이스에 존재하는 연관규칙의 발견에 대해서 주로 연구되어 왔으며, 정량적(수치적, quantitative) 속성을 갖는 데이터에 대한 연관규칙의 연구는 미비하였다. 그 이유는 정량적 속성을 갖는 데이터를 기호적(nominal) 속성값으로 바꿔주는 것이 어렵기 때문이다. 따라서 본 논문에서는 정량적 속성의 데이터를 기호적 속성값으로 바꾼 후 연관규칙을 추출함으로써, 퍼지 개념을 사용한 퍼지 연관규칙이 일반(범주, crisp, category) 개념을 사용한 연관규칙 보다 성능이 우수함을 보이고있다. 또한 본 논문에서는 퍼지 연관규칙에서 소속함수(항목, 아이템, 속성값)의 모양과 개수를 데이터 분포에 대한 통계적 특성을 나타내는 히스토그램을 이용하여 소속함수를 자동 생성하는 효율적인 연관규칙 추출방법을 제안한다.

1. 서론

현대 사회와 같이 데이터의 수집이 용이해진 지금 대량의 데이터에서 의미 있고 유용한 지식을 추출하는 일은 어려운 일이 되었다. 데이터에서 유용한 지식을 추출하는 과정을 데이터 마이닝(data mining)이라 한다. 데이터 마이닝은 사용되는 지식 추출 방법에 따라 얻을 수 있는 지식의 형태가 다양하다. 본 논문[1]에서는 마이닝 기법 중 연관규칙을 통해 유용한 지식을 추출하는 것을 목표로 한다. 본 논문에서 연관규칙을 통해 추출하는 유용한 지식이란 벤치마크 데이터를 높은 인식율(accuracy)로 분류할 수 있는 규칙을 말한다. 일반적으로 데이터는 레코드(트랜잭션)의 집합으로 되어있다. 또한 레코드는 속성과 속성값의 집합으로 구성되어있다. 연관규칙은 전체 데이터에서 특정 속성이 나타남과 동시에 다른 속성이 나타날 수 있는 빈도를 고려한 속성간의 연관 관계를 규칙으로 표현하게된다. 이렇게 추출된 연관규칙의 응용으로는 소비자 성향 분석을 예로 들 수 있다. 소비자들이 특정 품목을 구입하였을 때 어떤 다른 품목을 함께 구매한다는 규칙이 추출된다면 이 규칙을 통해 소비자의 구매 양식의 이해와 새로운 가치를 창출하는데 유용하게 사용될 수 있을 것이다. 그러나 기존에 연구되어온 연관규칙은 소비자가 어떠한 물품을 샀는지(이진 데이터)에 대한 구입 여부만을 규칙으로 생성하였는데, 그에 못지 않게 어떤 물품을 얼마만큼 구입하였는지(정량적 데이터)에 대한 데이터에서도 소비자의 성향을 분석하는데 좋은 규칙을 얻을수 있음을 알게되었다. 이처럼 정량적인 데이터에서 연관규칙을 추출하기 위해서는 정량적 데이터를 기호적 속성값으로 바꿔주는 일이 필요하다. 하지만 정량적 데이터를 기호적 속성값으로 바꿔주는 것은 어렵다. 예를 들어 속성 "온도"에 대한 속성값이 "25도" 라면 "25도"를 "온도가 중간이다"의 속성값으로 변환할 것인지, "온도가 높다"의 속성값으로 변환할 것인지에 따라 정량 데이터를 기호적 속성값으로의 변환이 오류 없이 잘 변환되었는지가 틀려지기 때문이다. 또한 규칙 추출 후 어떤 규칙으로 결론을 도출하느냐에 따라 규칙 적용이 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. 이때 실제 데이터의 적용에 있어서 두 규칙 중 좀 더 유용한 규칙을 적용하게 되는데 그 적용 방법에 따라 인식율의 차이가 생긴다. 이처럼 정량적인 데이터에 대한 데이터 변화와 규칙 적용 방법이 어렵다는 문제점을 해결하기 위해 퍼지 연관규칙과 연관규칙을 추출하여 어느 방법이 정량적인 데이터 처리에 더 유용한지를 평가하는 것이 본 논문의 내용이다. 본 논문의 구성은 다음과 같다. 2장에서는 퍼지 연관규칙 알고리즘과 연관규칙 알고리즘의 차이점에 대해 설명하고, 3장에서는 정량 데이터를 기호적 속성값으로 변환하는 방법과 퍼지 연관규칙에서 규칙 생성에 중요한 소속함수를 자동 생성하는 방법에 대해 설명하고, 4장에서는 1) 사용자가 소속함수를 주었을 때 퍼지 연관규칙 추출과 연관규칙 추출

1)본 연구는 한국과학기술원 뇌신경정보학 연구사업의 지원에 의하여 수행되었습니다.

인식율 평가, 2) 소속함수 자동생성하는 방법을 통해 퍼지 연관규칙 추출과 연관규칙 추출 인식율 평가에 대한 실험결과를 살펴보고, 5장에서 결론을 내린다.

2. 관련연구

논문에서 사용하는 연관규칙의 형태는 식(1)과 같고 연관규칙을 추출하기 위한 척도로는 지지도(support) 식(2)와 신뢰도(confidence) 식(3)이 사용된다. 지지도는 전체 데이터에서 속성에 대한 항목의 빈도수를 의미하고 사용자가 주게되는 최소 지지도 이상의 값만이 규칙을 생성할 수 있는 빈발 항목이 될 수 있다. 신뢰도는 규칙의 조건절 항목을 만족하는 데이터의 빈도에 대한 조건, 결론절 항목이 동시에 만족되는 빈도수를 의미한다.

규칙 : 만약 X가 A이면 Y는 B이다. (1)
(지지도 : 0.4, 신뢰도 : 1)

$$\text{지지도}\langle X,A\rangle=\frac{\sum_{t_i\in T}\prod_{x_j\in X}\{Ma_j\in A(t_i[x_j])\}}{T} \quad (2)$$

(T=전체레코드, i=전체레코드개수, X=속성들의집합, j=속성개수, A=소속함수들의집합, M=소속정도)

$$\text{신뢰도}\langle\langle X,A\rangle\langle Y,B\rangle\rangle=\frac{\sum_{t_i\in T}\prod_{z_k\in Z}\{Mc_k\in C(t_i[x_j])\}}{\sum_{t_i\in T}\prod_{x_j\in X}\{Ma_j\in A(t_i[x_j])\}} \quad (3)$$

($Z=X\cup Y$, $C=A\cup B$, k=속성개수)

정량 데이터를 이용하여 연관 규칙을 추출하는 방법으로는 퍼지 연관규칙과 연관규칙이 있다. 퍼지 연관규칙과 연관규칙의 차이점은 정량 데이터를 기호적 소속값으로 변환할 때 기호적 소속값의 소속정도(membership) 계산 방법과 추출된 여러 규칙에서 결론을 도출하기 위해 어떤 규칙을 적용하는지에 관한 추론(inference) 부분에서의 차이로 두 알고리즘을 설명할 수 있다.

2.1 소속정도 계산하는 방법

퍼지 연관규칙[1][2]은 정량 데이터의 속성값을 소속함수에 대한 소속정도로 변환될 때 소속정도가 0에서 1사이의 값을 갖는다. 반면에 연관규칙[3]은 소속정도가 0 또는 1의 이진 값을 갖는다. 퍼지 연관규칙에서 사용하는 소속함수의 모양과 연관규칙에서 사용하는 소속함수의 모양은 <그림1>과 같다. 퍼지 연관규칙에서 소속정도를 계산하는 방법은 식(6)(7)과 같다.

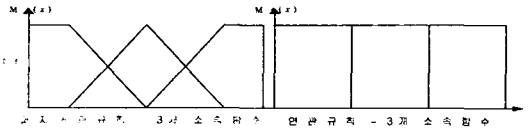

그림 1. 퍼지 연관규칙 소속함수와 연관규칙 소속함수의 형태

2.2 추론하는 방법

추론(inference)은 입력 데이터가 여러 규칙에 적용될 때 어떤 규칙을 이용하여 결론을 도출 할지에 관한 방법으로 퍼지 연관규칙은 식(4)에 의해 연관규칙은 식(5)에 적용하여 결론부를 적용할 규칙을 선택한다.

$$\text{확신도}_k(t_i)=\max_{R(l)}(O_1\{Ma_j\in A(t_i[x_j])\}\cdot O_2) \quad (4)$$

(R=규칙의집합, l=규칙개수, O_1= min 또는 *product*, O_2=지지도 * 신뢰도 또는 신뢰도)

$$\text{확신도}_k(t_i)=\max_{R(k)}O_2 \quad (5)$$

(R=규칙의집합, l=규칙개수, O_2=지지도*신뢰도 또는 신뢰도)

주어진 데이터에 대해서 각 클래스에 대한 확신도(Certainty Factor)는 식(4)(5)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 * 신뢰도 또는 규칙의 신뢰도)를 합성하는 연산
③ max: 각 규칙의 결과를 합성하는 연산

3. 제안하는 알고리즘

정량적인 데이터를 통해 효율적인 연관규칙을 추출하기 위해서는 사용자의 개입이 필수적이다. 사용자의 개입이란 규칙 추출을 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 <그림2>와 같이 삼각 소속함수와 사다리꼴 소속함수를 사용하고 식(6)(7)를 이용하여 소속정도를 구할 수 있다.

$$\text{삼각형}(\alpha,\beta,\gamma)=\begin{cases}0 & \text{if } x<\alpha\\ (x-\alpha)/(\beta-\alpha) & \text{if } \alpha\le x\le\beta\\ (\gamma-x)/(\gamma-\beta) & \text{if } \beta\le x\le\gamma\\ 0 & \text{if } x>\gamma\end{cases} \quad (6)$$

$$\text{사다리꼴}(\alpha,\beta,\gamma,\delta)=\begin{cases}0 & \text{if } x<\alpha\\ (x-\alpha)/(\beta-\alpha) & \text{if } \alpha\le x\le\beta\\ 1 & \text{if } \beta\le x\le\gamma\\ (x-\delta)/(\gamma-\delta) & \text{if } \gamma\le x\le\delta\\ 0 & \text{if } x>\delta\end{cases} \quad (7)$$

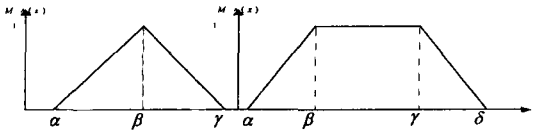

그림 2. 삼각 소속함수와 사다리꼴 소속함수

정량 데이터에서 추출된 연관규칙의 인식율을 평가하는 기존의 연구 방법으로는 사용자가 각 속성에 대하여 소속함수의 개수를 증가시킴으로써 인식율이 높이는 소속함수를 개수를 찾는 수동적인 방법을 사용하였다[4]. 가장 적절한 소속함수의 모양과 개수는 좋은 규칙을 생성하는데 가장 중요한 요소라고 볼 수 있다. 따라서 본 논문에서는 정량 연관규칙에서 가장 중요한 요소인 소속함수의 모양을 히스토그램(histogram)에 의해 자동 생성한다[5]. 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 극소

점을 직선으로 연결함으로 소속함수를 생성할 수 있다.

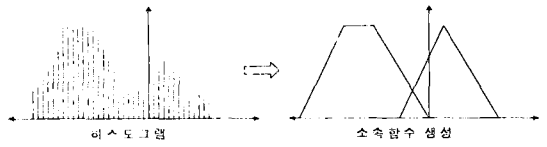

그림 3. 히스토그램 이용한 소속함수 생성

4. 실험

본 논문에서는 퍼지 연관 규칙과 연관 규칙의 성능을 평가하기 위해 패턴 분류 문제에 널리 사용되고 있는 벤치마크 데이터인 Iris 데이터와 Wine 데이터의 인식율을 비교하였다. Iris 데이터는 iris 꽃에 대해 3개의 품종(sectosa, versicolor, virginica)으로부터 각각 50개씩 총 150개의 개체를 추출하여 꽃받침 길이(sepal length), 꽃받침 폭(sepal width), 꽃잎 길이(petal length), 꽃잎 폭(petal width)을 센티미터(cm)단위로 측정한 데이터이고, Wine 데이터는 이탈리아에서 생산된 3개의 다른 품종의 포도주를 대상으로 알코올, 능금산, 마그네슘 등의 13개의 성분을 화학적으로 분석한 데이터로 총 178개의 개체를 가지고 있다[6]. 실험은 2-fold cross validation 방법을 통해 실험하였다. 실험에서 사용되는 규칙의 조건절은 iris 데이터의 경우 모든 속성이 조건절에 나오도록 규칙을 추출하였으며 반면 wine 데이터는 조건절의 속성의 개수를 2개로 한정하여 규칙을 추출하였다. 규칙의 결론절은 클래스 속성만을 갖도록 추출하였으며 규칙의 형태는 아래와 같다.

Iris 데이타 규칙:
만약 꽃받침길이가 짧고 꽃받침폭이 중간이고
꽃잎길이가 짧고 꽃잎폭이 좁으면 클래스는 sec*tosa*이다.
(지지도 = 0.22, 신뢰도 = 1)

Wine 데이타 규칙:
만약 능금산이 적고 프롤린이 많다면 클래스는 0이다.
(지지도 = 0.16, 신뢰도 = 0.99)

규칙 추출을 위해 최소 지지도는 0, 최소 신뢰도는 0.9값을 사용하였다. <표1>, <표2>, <표3>에서 확신도①과②는 확신도 계산 후 규칙 적용에 대한 두가지 추론 방법으로 ①은 지지도 * 신뢰도의 최대값을 ②는 신뢰도의 최대값을 갖는 규칙을 이용하여 결론을 도출하였다.

(1) 각 항목(소속함수)별 퍼지 연관규칙과 연관규칙 인식율 비교

표 1. Iris 데이터 항목별 인식율(%)

확신도	적용	2 항목	3 항목	4 항목	5 항목	평균
①	퍼지	54	88	96.6	94.6	83.3
	일반	35.3	96	75.3	88	73.65
②	퍼지	63.3	94.6	96.6	96	87.62
	일반	35.3	96	78	88.6	74.47

표 2. Wine 데이터 항목별 인식율(%)

확신도	적용	2 항목	3 항목	4 항목	5 항목	평균
①	퍼지	79.2	90.4	85.3	93.2	89.52
	일반	66.8	81.4	84.8	89.8	80.7
②	퍼지	80.8	91.5	92.6	93.8	89.67
	일반	71.3	81.4	92.6	87	83.07

(2) 히스토그램 이용 소속함수 생성 인식율(%)

표 3. 히스토그램 이용 소속함수 생성 인식율(%)

확신도	적용	Iris data	Wine data
①	퍼지	95.3	84.8
	일반	92.6	83.7
②	퍼지	94.6	85.3
	일반	92.6	84.2

5. 결론 및 향후 연구

본 논문에서는 벤치마크 데이터를 이용하여 퍼지 연관규칙과 연관규칙 추출을 통해 인식율를 비교함으로 퍼지 연관규칙이 연관규칙보다 성능이 우수함을 보였다. 퍼지는 언어적 불확실성을 퍼지 소속함수의 개념을 이용하여 수치적으로 표현하여 잡음이 있거나 불확실한 데이터의 처리에 유용하며, 퍼지 연관규칙을 사용할 경우 추론의 효율성을 크게 높일 수 있으며, 이해하기 쉽고 인간의 지식과 잘 통합될 수 있는 규칙을 추출할 수 있는 장점을 갖는다. 향후연구계획으로는 제안한 알고리즘을 실제 데이터에 적용하여 보다 일반적인 타당성을 검증하고, 히스토그램 기반 소속함수의 조율을 통해 보다 높은 인식율을 얻을 수 있는 소속함수를 생성한다.

6. 참고 문헌

[1] Chan Man Kuok, Ada Fu, Man Hon Wong, "Mining Fuzzy Association Rules in Databases", SIGMOD record, Vol. 27, No 1, pp. 41-46, 1998

[2] Attila Gyenesei, "A Fuzzy Approach for Mining Quantitative Association Rules", TUCS Technical Reports, No336, 2000

[3] Ramakrishnan Srikant, Rakesh Agrawal, "Mining Quantitative Association Rules in Large Relational Tables", Proceedings of the 1996 {ACM} {SIGMOD} International Conference on Management of Data, pp. 1-12, 1996

[4] Hisao Ishibuchi, Tomoharu Nakashima and Takashi Yamamoto, "Fuzzy association rules for handling continuous attributes", IEEE International Symposium, Vol. 1, pp. 118-121, 2001

[5] Myung Won Kim, Joong Geun Lee and Changwoo Min, "Efficient Fuzzy Rule Generation Based on Fuzzy Decision Tree", IEEE International Fuzzy Systems Conference Proceedings, pp.1223-1228, 1999

[6] ftp://ftp.ics.uci.edu/pub/machine-learning-databases/