퍼지 개념 계층을 도입한 일반화된 연관 규칙 마이닝

손봉기 김동호 이건명
충북대학교 컴퓨터 과학과, 첨단정보기술 연구센터

Mining Generalized Association Rules Using Fuzzy Concept Hierarchy

Bong-Ki Sohn, Dong-Ho Kim, Keon-Myung Lee
Dept. of Computer Science, Chungbuk National University, and AITrc

요 약

연관 규칙 마이닝 과정에 참조되는 일반 개념 계층은 개념 간의 명확한 관계만을 표현한다. 실제로는 개념 사이의 관계가 애매한 경우가 많다. 이 논문에서는 개념간의 애매한 관계까지 반영할 수 있는 퍼지 개념 계층을 이용하여 일반화된 연관 규칙을 마이닝하는 방법을 제안한다. 퍼지 개념 계층에서의 하위 개념을 상위 개념으로 적절하게 반영하는 방법과 마이닝된 연관 규칙에서 중복되는 규칙의 가지치기(pruning)에 사용되는 측도를 소개한다. 또한 퍼지 개념 계층을 이용한 일반화된 연관 규칙 마이닝 방법의 응용성을 보이기 위해 실험 과정과 결과를 보인다.

1. 서 론

연관 규칙 마이닝(mining association rule)은 많은 트랜잭션 항목(transaction items)들에 대해 "트랜잭션에서 몇 가지 항목이 나타나면 같은 트랜잭션에 다른 항목들이 나타난다"와 같이 항목들 사이의 중요한 연관성을 찾아내는 것이다[1]. 연관성 정도는 지지도(support degree)와 신뢰도(confidence degree)로 측정한다. 지지도는 전체 트랜잭션 개수 중 해당 연관 규칙이 지지하는 트랜잭션의 비율을 나타내며, 최소 지지도(minimum support) 이상의 항목들을 식별하는데 사용된다. 신뢰도는 연관 규칙의 강도(strength)를 의미하며, 전제부(antecedent)를 만족하는 트랜잭션이 결론부(consequent)까지를 만족하는 비율로서, 최소 신뢰도(minimum confidence) 이상의 연관 규칙을 도출하는데 사용된다.

연관 규칙 마이닝 과정에는 항목들 사이의 개념 계층(taxonomy)을 도입함으로써 유용한 경우가 많다[2][3]. 이러한 방법들은 개념 계층을 도입하지 않은 연관 규칙 마이닝 방법에 비해[4,5,6,7] 다음과 같은 장점들을 갖는다. 첫째, 개념 계층의 낮은 수준에서의 규칙들은 최소 지지도를 갖지 않는 경우가 많기 때문에 중요한 연관 규칙을 상위 수준에서 발견할 수 있다. 둘째, 개념 계층이 흥미롭지 않거나 중복되는 규칙(redundant rules)을 가지치기(pruning)하는데 사용될 수 있다. 따라서 연관 규칙을 마이닝하는데 참조되는 개념 계층은 개념들 사이의 관계가 적절하게 표현되어야 한다. 그러나 실제로 개념들 사이의 관계는 명확하지 않고 애매한 경우가 많다. 기존의 연구에서는 개념들 사이의 관계를 명확하게 표현하였지만, 애매한 관계를 개념 계층에 반영하려는 시도가 없었다.

본 연구는 첨단정보기술센터(AITrc)를 통해 과학재단 지원으로 수행된 것임

이 논문에서는 개념들 사이의 일반화 정도를 반영하는 퍼지 개념 계층을 참조하여 일반화된 연관 규칙을 마이닝하는 알고리즘을 제안한다.

이 논문은 다음과 같이 구성되어 있다. 2절에서는 일반화된 연관 규칙을 마이닝하는데 참조되는 퍼지 개념 계층에 대해 기술한다. 3절에서는 퍼지 개념 계층을 도입한 일반화된 연관 규칙 마이닝 알고리즘을 제안하고, 4절에서는 제안된 방법으로 퍼지 개념 계층을 이용해 일반화된 연관 규칙을 도출하는 과정을 예를 통해 알아본다. 5절에서는 결론과 향후 연구 과제를 제시한다.

2. 퍼지 개념 계층

개념 계층(concept hierarchy)은 영역 개념들 사이의 일반화 관계를 표현하는데 사용된다. 일반 개념 계층은 개념들 사이의 명백한 일반화 관계를 나타낸다. 개념 계층은 비순환 유향 그래프 (N,A)로 표현되는데,N은 개념 노드의 집합이고 A는 일반화 관계를 나타내는 간선 (n_i, n_j)의 집합이다. 간선 (n_i, n_j)는 n_i가 n_j의 일반화 개념이라는 것을 의미한다. 일반 개념 계층에서 일반화 관계를 나타내는 모든 간선들은 명백하다.

퍼지 개념 계층은 퍼지 간선 $(n_i, n_j, \gamma_{n_i n_j})$로 표현되는데, $\gamma_{n_i n_j}$는 n_i의 n_j로의 일반화 정도이다. 퍼지 간선 $(n_i, n_j, \gamma_{n_i n_j})$의 n_j는 $\gamma_{n_i n_j}$정도로 n_j를 부분적으로 일반화한 개념이라는 것을 의미한다. 퍼지 개념 계층에서의 개념은 부분적으로 여러 개의 일반화 개념으로 일반화되는 관계를 가질 수 있다. 퍼지 개념 계층은 애매한 개념들 사이의 일반화 관계를 표현하는데 적절하기 때문에, 트랜잭션의 각 항목들이 상위 개념으로 제대로 반영되어 의미 있는

연관 규칙의 마이닝이 가능하다. (그림 1)은 퍼지 개념 계층의 예를 보인 것인데, 간선에 기입된 숫자는 해당 개념들 사이의 일반화 정도를 나타낸다. 예를 들어, G는 1정도로 D에 속하고, 0.7정도로 E에 속한다는 것을 의미한다.

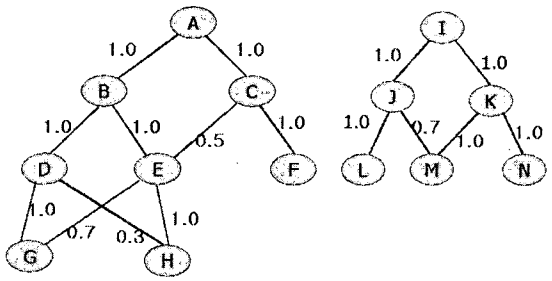

그림 1. 퍼지 개념 계층 집합

3. 퍼지 개념 계층을 도입한 일반화된 연관 규칙 마이닝

이 논문에서 제안한 퍼지 개념 계층을 이용한 일반화된 연관 규칙 마이닝 방법에서 사용하는 몇 가지 개념을 다음과 같이 정의한다.

- 일반화 정도 γ_{ab}
  개념 a가 개념 b로 일반화되는 정도
- 일반화 정도 한계값 θ_g
  개념 a가 상위 개념 b로 일반화되는 최소 일반화 정도
- 포함 정도(coverage degree) c_a
  트랜잭션의 항목 a가 상위 개념으로 표현되는 정도
- 만족 정도(satisfaction degree) S
  frequent itemset 또는 연관 규칙의 구성 개념들의 평균 포함 정도의 최소값
- 최소 만족 정도 θ_s
  frequent itemset 또는 연관 규칙이 가져야할 최소 만족 정도
- 흥미도(interestingness degree) I_a
  마이닝된 연관 규칙의 유효성을 나타내는 측도
- 최소 흥미도 θ_i
  유효한 규칙이 되기 위한 최소 흥미도
- 확장 트랜잭션
  트랜잭션의 모든 항목들과 각 항목들의 일반화 한계값 이상의 모든 상위 개념으로 이루어지고, 중복되는 개념이 합병된 트랜잭션

이 논문에서 제안한 연관 규칙 마이닝 과정은 주어진 퍼지 개념 계층과 트랜잭션 데이터베이스로부터 확장 트랜잭션 데이터베이스를 생성하고, 이로부터 frequent itemset을 구한다. frequent itemset에서 최소 신뢰도 이상인 연관 규칙을 추출 최소 흥미도를 이용하여 최종 연관 규칙을 도출한다.

3.1 확장 트랜잭션의 생성

주어진 데이터베이스에서 트랜잭션을 확장 트랜잭션으로의 변경은 트랜잭션의 모든 항목들에 일반화 한계값 이상인 각 항목들의 상위 개념들을 퍼지 개념 계층을 참조하여 추가함으로써 이루어진다. 확장 트랜잭션은 (개념,포함 정도)의 집합으로 표현된다. 예를 들어,(그림 1)에서 H의 상위 개념 B를 확장 트랜잭션에 표현할 때 B의 포함 정도는 t-norm함수를 이용해 결정한다. 즉, $c_B = ①(c_E, \gamma_{EB})$와 $c_B = ①(c_D, \gamma_{DB})$로 표현된다. 각 개념들은 여러 상위 개념으로 일반화될 수 있기 때문에 한 트랜잭션에서 포함 정도는 다르지만 동일한 개념이 포함될 수 있다. 중복되는 개념들은 t-conorm함수에 의해 하나의 개념으로 합병된다. H의 상위 개념인 B는 포함 정도는 다르지만 동일한 개념 B가 두 번 같은 트랜잭션에 나타나는데, 하나의 개념 B와 포함 정도는 $(B, ⑤(c_B = ①(c_E, \gamma_{EB}), c_B = ①(c_D, \gamma_{DB}))$로 합병한다. 이러한 과정을 통해 확장 트랜잭션을 생성한 후 이 확장 트랜잭션 데이터베이스를 기반으로 하여 frequent itemset을 구한다.

3.2 Frequent-itemset 추출 알고리즘

다음은 확장 트랜잭션 데이터베이스를 통해 frequent itemset을 추출하는 알고리즘이다.

Procedure fuzzy-frequent-itemsets
입력
확장 트랜잭션 데이터베이스 D
퍼지 개념 계층 집합 T
최소 지지도 min-sup
만족 정도(satisfaction degree) S
최소 만족 정도 θ_s
출력
min-sup이상의 모든 frequent itemsets
begin
L_1 := {frequent 1-itemsets};
L_1 := L_1 - (L_1의 각 frequent itemset의 S ≤ θ_s)
k := 2;
While($L_{k-1} \neq 0$) **do**
begin
C_k := L_{k-1}로부터 생성된 크기 k의 새로운 후보
C_k := C_k - 개념과 그 개념의 상위 개념으로 이루어진 모든 후보
forall transaction t∈ D **do**
begin
C_k에 있는 모든 후보에 대해 횟수 증가
C_k에 있는 모든 후보에 대해 S 계산
end
C_k := C_k - { C_k중 θ_s보다 작은 S를 갖는 후보}
L_1 := min-sup를 갖는 C_k의 모든 후보
K := $k+1$
end
Answer := $\bigcup_k L_k$

3.3 연관 규칙의 추출과 가지치기

연관 규칙은 frequent itemset으로부터 [5]의 방법에 의해 최소 신뢰도 이상인 규칙만을 추출한다. 이 논문에서 제안한 연관 규칙 마이닝 방법에서는 최소 지지도와 최소 신뢰도 뿐만 아니라 최소 만족 정도를 적용하여 frequent itemset의 후보를 생성할 때, 만족 정도가 낮은 후보를 제외시킨다. 또한 이러한 측도 이외에 추출된 연관 규칙에 대한 유효성(usefulness)을 측정하기 위해 흥미도를 이용한다. 흥미도를 이용함으로써, 발견된 규칙이 다른 규칙에 비해 얼마나 유용한가를 알 수 있고 최소 흥미도 이상인 규칙만을 최종 연관 규칙으로 결정하게 된다. 이 논문에서 사용하는 흥미도에 대한 측도는 다음과 같다.

$$E_{\hat{Z}}[\Pr(Z)] = \frac{\Pr(z_1) \times \gamma_{z_1 \hat{z}_1}}{\Pr(\hat{z}_1)} \times \ldots \times \frac{\Pr(z_j) \times \gamma_{z_j \hat{z}_j}}{\Pr(\hat{z}_j)} \times \Pr(\hat{Z})$$

즉, $\hat{Z}$가 Z의 상위개념이고, $Pr(\hat{Z})$이 주어질 때 $Pr(Z)$의 기대값을 구해봄으로서 발견된 연관 규칙이 유효한 지를 검사할 수 있다. 여기서 연관 규칙의 각 항목들 간의 일반화 정도를 반영함으로서 기대값을 더 정확하게 예측할 수 있다.

4. 실 험

퍼지 개념 계층을 도입한 연관 규칙 마이닝 방법의 응용성을 보이기 위해 이 절에서는 제안한 방법을 적용한 예를 보인다. 퍼지 개념 계층과 트랜잭션 데이터베이스가 (그림 1)과 (표 1)로 주어지고, 최소 지지도를 33.3%(2 트랜잭션),일반화 정도 한계값 θ_g를 0.5, 최소 만족 정도 θ_s를 0.75로 하였다.

표 1. 트랜잭션 데이터베이스

TID	Items
100	{F}
200	{G,N}
300	{H,M}
400	{F,L}
500	{G}
600	{F,L,N}

(표 2)는 퍼지 개념 계층을 참조해 트랜잭션의 각 항목들의 상위 개념과 포함정도를 나타낸 확장 트랜잭션 데이터베이스를 나타낸 것이다.

표 2. 확장 트랜잭션 데이터베이스

TID	Items
100	{ (A,1.0),(C,1.0),(F,1.0) }
200	{(A,1.0),(B,1.0),(C,0.5),(D,1.0),(E,0.7), (G,1.0),(I,1.0),(K,1.0) (N,1.0) }
300	{(A,1.0),(B,1.0),(C,0.5),(E,1.0),(H,1.0), (I,1.0),(J,0.7), (K,1.0),(M,1.0) }
400	{ (A,1.0),(C,1.0),(F,1.0),(I,1.0),(J,1.0),(L,1.0) }
500	{ (A,1.0),(B,1.0),(C,0.5),(D,1.0),(E,0.7),(G,1.0)}
600	{(A,1.0),(C,1.0),(F,1.0),(I,1.0),(J,1.0), (K,1.0),(L,1.0), (N,1.0) }

확장 트랜잭션 데이터베이스로부터 최소 지지도와 최소 만족 정도이상의 frequent itemset을 추출하고 (표 3)와 같이 최소 신뢰도를 만족하는 규칙을 도출해 낸다.

(표 4)는 (표 3)으로부터 최소 흥미도 θ_i = 1.5를 만족하는 규칙들로서 최종 연관 규칙들이다. 즉, (표 1)의 트랜잭션들로부터 퍼지 개념 계층을 참조하여 일반화된 연관 규칙을 도출한 결과이다. 각각의 규칙은 신뢰도와 만족 정도로 규칙을 기술할 수 있다. 예를 들어, 규칙 A->I는 "100% A품목에 속하는 상품을 구매하는 사람의 66.6%가 100% I품목에 속하는 상품도 구매한다"라고 해석할 수 있다.

5. 결 론

연관 규칙 마이닝에 대한 많은 연구가 있었지만, 개념들 사이의 애매한 관계를 표현할 수 있는 퍼지 개념 계층을 적용한 연구는 많이 시도되지 않았다. 이 논문에서는 퍼지 개념 계층을 이용해 효과적으로 연관 규칙을 마이닝할 수 있는 방법을 제안하였다. 또한 실험을 통해 퍼지 개념 계층을 이용한 연관 규칙 마이닝이 의미있는 규칙을 추출한다는 것을 확인할 수 있었다. 퍼지 개념 계층에서의 개념은 부분적으로 여러 개의 상위 개념으로 일반화되는 관계를 가질 수

표 3. 최소 신뢰도를 만족하는 연관 규칙

Rules	support	confidence	만족정도
A->I	4	66.6%	{1.0}
B->I	2	66.6%	{1.0}
B->K	2	66.6%	{1.0}
C->I	4	66.6%	{0.75}
E->I	2	66.6%	{0.85}
E->K	2	66.6%	{0.85}
F->I	2	66.6%	{1.0}
F->J	2	66.6%	{1.0}
F->L	2	66.6%	{1.0}
J->K	2	66.6%	{0.85}
JK->A	2	100%	{0.85}
AK->J	2	66.6%	{0.85}
AJ->K	2	66.6%	{0.85}
J->AK	2	66.6%	{0.85}
K->AJ	2	66.6%	{0.85}

표 4. 최종 연관 규칙

Rules	support	confidence	만족정도
A->I	4	66.6%	{1.0}
E->I	2	66.6%	{0.85}
E->K	2	66.6%	{0.85}
F->I	2	66.6%	{1.0}
F->L	2	66.6%	{1.0}
J->K	2	66.6%	{0.85}
JK->A	2	100%	{0.85}
AK->J	2	66.6%	{0.85}
AJ->K	2	66.6%	{0.85}
J->AK	2	66.6%	{0.85}
K->AJ	2	66.6%	{0.85}

있다. 이는 방대한 트랜잭션 데이터베이스를 대상으로 하는 연관 규칙 마이닝에 있어 저장장소와 효율성 문제를 야기할 수 있다는 것을 의미한다. 또한 마이닝된 연관 규칙 에 중복된 규칙이 많다는 것을 의미한다. 따라서 이러한 효율성 문제와 마이닝된 규칙에 대한 가지치기 방법에 대한 추가 연구가 필요하다.

참고 문헌

[1] Ming-Syan Chen, Jiawei Han, Philip S.Yu, Data Mining:An overview from Database Perspective. IEEE Trans.on Knowledge and Data Engineering, 1997.
[2]R. Srikant, R. Agrawal: "Mining Generalized Association Rules", Proc. of the 21st Int'l Conference on Very Large Databases, Zurich, Switzerland, Sep. 1995.
[3] Jiawei Han, Yongjian Fu, Discovery of Multiple-level Association Rules from Large Databases
[4]Rakesh Agrawal, Tomasz Imielinski, and Arun Swami. Mining association rules between sets of items in large databases. In proc. of the ACM SIGMOD Conference on Management of Data, pages 207-216, Washington, D.C., May 1993.
[5] Rakesh Agrawal and Ramakrishnan Srikant. Fast algorithms for mining association rules. In proc. of the VLDB Conference, Santiago, Chile, September 1994. Expanded version available as IBM Research Report RJ9839, June 1994.
[6] Maurice Houtsma and Arun Swami. Set-oriented mining of association rules. In Intl Conference on Data Engineering, Taipei, Taiwan, March 1995.
[7] Heikki Mannila, Hannu Toivonen, and A.Inkeri Verkamo. Efficient algorithms for discovering association rules. In KDD-94: AAAI Workshop on Knowledge Discovery in Databases, pages 181-192, Seattle, Washington, July 1994.