복수 샘플링과 트리밍을 통한 고품질 연관규칙 추출법

황원태 김동승
고려대학교 전기공학과
wthwang@classic.korea.ac.kr dkim@classic.korea.ac.kr

Improved Association Rule Mining by Multiple Sampling & Trimming

Wontae Hwang Dongseung Kim
Dept. of Electrical Engineering, Korea University, Seoul, Korea

요 약

본 논문은 전체 데이터베이스에서 일부 추출된 샘플 데이터에서 빈발항목 집합을 찾는 연관규칙 마이닝 알고리즘을 기술한다. 샘플링기술을 이용하면 마이닝과정에서 필요한 데이터베이스의 접근 양을 줄이므로써 실행시간을 단축시킬 수 있다는 장점이 있지만, 전체데이터베이스를 이용한 마이닝보다 정확도가 떨어진다는 단점이 함께 존재한다. 이전의 Chen의 FAST알고리즘은 샘플링을 이용한 마이닝과정에서 거리오차함수를 이용한 트리밍과정을 통해 빈발 1항목집합에 대한 정확도를 개선시켰다. 이후 IFAST 알고리즘은 트리밍과정에서 빈발2-항목집합까지 고려하여 빈발2-항목집합 이상의 빈발항목집합에서도 정확도를 개선시켰다. 본 논문에서는 트리밍과정에서 사용될 추정데이터를 여러 개의 샘플데이터를 이용하여 얻으므로써 오류항목집합(false itemset)의 수를 줄이고 전체적인 정확도를 향상시키는 새로운 알고리즘을 소개한다.

1. 서 론

연관규칙 마이닝은 대규모 자료(Database)를 분석하고 그 속에서 의미 있는 패턴을 추출하는 데이터마이닝의 한 분야이다. 연관규칙은 트랜잭션(transaction) 데이터로부터 추출되며, 슈퍼마켓에서 고객들의 물건 구입패턴을 찾는 것을 예로 들 수 있다[1]. 이때 고객 한명은 하나의 트랜잭션이 되고 슈퍼마켓의 물품들은 항목(item)이 되며 여기서 연관 규칙 마이닝은 구입물품자료에서 물품들 간의 강한 상관성을 찾는 과정이다. 만약 데이터베이스를 마이닝한 결과 분유를 사는 고객이 기저귀를 사거나 반대로 기저귀를 사는 고객이 분유를 살 확률이 높다는 정보를 추출해 내었다면 이점을 이용 기저귀와 분유를 가까운 위치에 배치하여 고객편의를 돕거나, 나아가 매출신장에 이용할 수 있을 것이다. 이러한 연관 규칙 마이닝은 슈퍼마켓뿐만 아니라 신용카드회사의 카드불법사용감지, 통신회사의 이동전화불법사용감지, 전자상거래, 의사결정 지원 및 의료분야 등에도 활용될 수 있다.

연관 규칙을 추출하기 위해서는 지지도(support)와 신뢰도(confidence) 두 가지 설정 값이 있어야 한다. 연관 규칙 추출에 있어 지지도와 신뢰도의 정의는 표 1과 같다. 여기서 I는 항목집합, D는 전체 데이터베이스 그리고 T_i는 i번째 트랜잭션 데이터를 뜻한다.

연관규칙 추출 알고리즘은 주어진 최소지지도와 최소신뢰도 값 이상의 모든 연관규칙을 찾아낸다. 연관 규칙 마이닝은 다음의 두 단계로 구성된다. 첫번째 단계는 데이터베이스 D에서 최소지지도 이상의 지지도를 갖는 모든 빈발 항목집합들을 찾는 과정이고, 다음 단계는 찾아진 항목집합 중 최소신뢰도를 만족하는 빈발항목집합을 찾는 과정이다. 이때 두번째 단계는

표 1. 연관 규칙

$I=\{i_1, i_2, \ldots, i_m\}$, $D=\{T_1, T_2, \ldots, T_n\}$, $T_i \subset I$,
$X, Y \subset I$ 그리고 $X \cap Y = \varnothing$ 일때
Association rule(R): X -> Y
support(X) = X의 출현 횟수 / |D|
confidence(R) = support(X∪Y) / support(X)

비교적 간단하고 계산비용이 덜 드는 과정으로, 연관 규칙 추출에 관한 대부분의 연구는 첫번째 단계인 빈발 항목집합을 효율적으로 찾는 과정에 집중되고 있다. 본 논문 또한 빈발 항목집합을 신속하게 찾는 방법으로 샘플링 기반의 알고리즘을 제안한다.

2. IFAST

연관규칙 마이닝의 대표적인 Apriori[2]알고리즘은 '빈발하지 않는 집합의 superset은 빈발하지 않다.'는 원칙을 적용하여 빈발할 가능성이 있는 후보 항목집합의 수를 줄임으로써 후보 항목집합의 생성에 드는 비용과 그것의 지지도를 계산하는 비용을 상당히 줄였다. 그럼에도 불구하고 Apriori 유사 알고리즘은 후보자 집합을 구하는데 많은 계산량을 요구하고 후보자 집합들을 검사하여 빈발항목집합을 찾는 과정에서 데이터베이스를 여러 번 스캔(scan)함으로써 많은 시간을 소요하는 단점이 있었다.

이후 그러한 단점을 개선시켜 데이터베이스 스캔수를 줄이거나 후보자집합을 생성하지 않고 빈발항목집합을 보다 효율적으로 찾는 Partition[3], FP-Growth[4], COFI-tree[5] 알고리즘 등이 소개되어왔다.

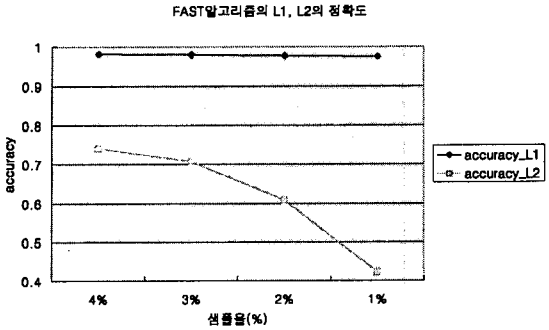

그림 1. FAST 알고리즘의 L_1, L_2의 정확도 (|D|=100,000)

한편 방대한 양의 데이터베이스 크기 자체를 줄인 샘플 데이터를 이용하여 소요시간을 단축하는 샘플링기반의 알고리즘들이 소개되기도 하였다. 샘플링기법을 이용한 접근 방법은 상대적으로 소규모의 데이터를 이용하므로 실행시간을 줄일 수 있는 반면 Toivonen의 논문[6]에서도 확인되었듯이 실제 데이터베이스의 일부에서 빈발항목집합을 찾을 경우 미발견(missing) 항목집합과 오류(false) 항목집합이 발생하여 빈발항목집합의 정확도가 낮아지는 문제점이 있다(미발견 항목집합은 전체 데이터에서 보면 빈발한 항목인데 샘플 데이터에서는 빈발하지 않아 제외된 데이터 집합이고, 오류 항목집합은 전체 데이터에서는 빈발하지 않는데도 불구하고 샘플 데이터에서는 빈발한 것으로 판단되어 뽑힌 항목이다). 이에 따라, 샘플링기반의 알고리즘은 실행시간과 정확도간의 trade-off가 존재한다. 이에 대해 최근 Chen은 실행시간과 정확도의 상관관계에서 이전의 샘플링 기반의 연관 규칙 추출 알고리즘들 보다 좋은 정확도를 보이는 2단계 샘플링 FAST 알고리즘[7]을 제안하였다. 하지만 이전연구에서 보고한 바와 같이 실험을 통해 FAST 알고리즘이 빈발 1-항목 집합에 대해서는 정확도가 높지만 2-항목 집합이상에 대해서는 정확도가 떨어지는 단점(그림 1)이 있음을 알아냈고 그것을 개선하는 IFAST 알고리즘을 제시하였다[8]. 본 논문은 IFAST를 더욱 개선하고자 시도된 것으로 빈발 2-항목집합이상의 항목집합에 대해서 정확도를 이전보다 높일 수 있는 복수 샘플링 방식의 알고리즘을 소개한다.

우선 설명을 쉽게 하기 위해 우리는 연관규칙 마이닝에 대한 논문에 적절한 용어를 표 2에 정리하였다.

표 2. Notation

D: 데이터베이스 (D={T_1, T_2, ..., T_n})
S: D로부터 교체 없이 뽑은 단순 랜덤 샘플
I: D에 나타나는 모든 항목들의 집합
N(=|D|), n(=|S|): 각각 D와 S내의 트랜잭션의 수
m(=|I|): 항목들의 수
I(D): D내에 나타나는 항목들의 총 집합
I_k(D), I_k(S): 각각 D와 S내의 k-항목집합들의 모음
n(A; T): 항목집합A를 포함하는 T내의 트랜잭션의 수
D내의 A의 지지도: f(A; D)=n(A; D)/|D|
S내의 A의 지지도: f(A; S)=n(A; S)/|S|
L(D), L(S): 각각 D와 S내에서 빈발 항목집합
L_k(D), L_k(S): 각각 D와 S내에서의 빈발 k-항목집합

2.1 IFAST 알고리즘

FAST와 IFAST 알고리즘에서는 1-항목집합 사이의 정확한 차이의 정도를 나타내기 위해 [7,8]에서는 다음과 같은 거리오차 함수 *Dist*를 사용하고 있다.

$$Dist_1(S_0,S)=\frac{|L_1(S)-L_1(S_0)|+|L_1(S_0)-L_1(S)|}{|L_1(S_0)|+|L_1(S)|}$$

이 함수는 0에서 1사이의 값을 갖으며 0에 가까울수록 S_0와 S는 같은 집합이 된다. 그리고 이 함수는 오류 빈발1-항목집합과 미발견 빈발 1-항목집합 과다에 따라 민감하게 변한다. [8]에서는 또한 2-항목집합에 대한 고려를 하기 위해 다음과 같은 $Dist_{L2}$ 함수를 사용한다.

$$Dist_{L2}(F,t^*)=\frac{|L_2(F)-L''_2(t^*)|}{|L_2(F)|}$$

$Dist_{L2}$ 함수는 트랜잭션을 구성하고 있는 항목 중에서 L_2(F)에 포함된 2-항목집합이 많을수록 함수 값이 작게 되므로 큰 함수 값을 갖는 트랜잭션을 찾아 이질자(outlier)로 판별하게 된다. 지면관계상 이질자 제거과정과 정확도측정에 관한 자세한 내용은 [8]을 참고하기 바란다.

IFAST(Improved Finding Associations from Sampled Transactions) 알고리즘은 임의 추출된 샘플데이터 가운데서 빈발 1-항목집합은 물론 빈발 2-항목집합의 출현정도를 동시에 고려하여 최종샘플데이터를 구하도록한 알고리즘이다. [8]에서는 IFAST 알고리즘이 FAST 알고리즘에 비해 동일한 실행시간에 보다 높은 정확도의 결과를 얻음을 확인하였다. 하지만 IFAST 알고리즘은 trimming과정에서 원래의 빈발 1-항목집합 대신 샘플에 의해 추정된 빈발 2-항목집합을 사용함으로써 예상과는 달리 오류항목집합의 수를 증가 시킬 수 있다. 이러한 사실이 확인되면 최악의 경우 알고리즘을 처음부터 다시 실행해야할지도 모른다는 문제점을 가지고 있다.

3. 새로운 알고리즘

3.1 복수 샘플링 알고리즘

실험을 통해 샘플정도를 크게 할수록 IFAST 알고리즘 과정 후의 샘플데이터에서 찾아진 오류 항목집합의 수가 미발견 항목집합에 비해 상대적으로 크게 증가함을 볼 수 있었다. 오류 항목집합의 수를 줄이기 위해 복수 샘플링 알고리즘에서는 trimming과정에서 이용되는 항목들을 결정함에 있어 복수의 샘플세트를 이용하고 각각의 세트에서 도출된 결과를 다수결 방식에 의해 최종적 코어셋을 찾게 되어 보다 신뢰도를 높일 것을 예상한다. 복수 샘플링방식의 구체적인 알고리즘은 표 3과 같다.

실행과정을 설명하면, 우선 복수 샘플링 알고리즘은 전체 데이터 D에서 단순 랜덤 샘플 S를 구한다. 그리고 S를 Q개(Q는 홀수)의 동일한 트랜잭션을 가지지 않는 샘플로 나눈다. Q개의 샘플에서 다음 trimming 과정에 이용될 추정데이터(샘플의 일부에서 추출된 빈발 1-항목집합(L1')과 빈발 2-항목집합(L2'))를 각각 구한다. 구해진 추정데이터와 IFAST 알고리즘의 아웃라이어 제거방법을 이용해 Q개의 샘플마다 데이터 C_i(i=1, ..., Q)를 구한다. 이때 C_i의 크기는 사용자가 정해놓은 최종샘플의 크기와 같다. 다음단계에서는 구해진 C_i의 추정데이터를 구해 최종샘플을 구하는 trimming 과정에 이용될 추정데이터를 결정

표 3. 복수 샘플링 알고리즘

Multiple_IFAST (D, Q, n_c=|S_0|)
1. D로부터 simple random sample S를 구한다.;
2. S를 Q개의 disjoint sample sets(S_i; i=1, ..., Q)로 나눈다.;
3. Q개의 sample sets에서 trimming과정에서 이용될 추정데이터(추정빈발항목집합 L1', L2')를 각각 구한다.
4. Q개의 sample set의 각각에서 3에서 구해진 추정데이터를 각각 사용하여 trimming을 진행하여 Q개의 데이터 C_i(|C_i|=n_c)를 구한다.
5. 4에서 구해진 각 C_i(i=1, ..., Q)의 추정데이터(추정빈발 항목집합 L1', L2')를 얻어 6의 최종trimming과정에 이용될 새로운 추정데이터 ((L_1^T)', (L_2^T)')를 다음과 같이 선택한다.
   어느 하나의 추정데이터가 Q개 C_i의 추정데이터에서 나타난 횟수가 나타나지 않은 횟수보다 많으면 다수결 방식에 의해 그 데이터를 새 추정데이터 ((L_1^T)', (L_2^T)')로 결정.
6. 각 C_i(i=1, ..., Q)의 합집합($C_1 \cup C_2 ... \cup C_Q$)을 5에서 구해진 추정데이터(($L_1^T$)', ($L_2^T$)')를 가지고 trimming하여 최종샘플 S_0(|S_0|=n_c)를 구한다.
7. return S_0;

하게 된다. 이전 단계까지의 추정데이터로는 IFAST 에서와 동일하게 데이터의 일부에서 추출되어진 빈발항목집합을 사용하였지만 이번 단계에서는 Q개의 추정데이터에서 다수결 원칙에 의해 선택된 데이터들을 새로운 추정데이터로 사용한다. 이렇게 trimming 과정에서 쓸 추정데이터를 신중하게 얻어서 오류항목집합, 즉 잘못 찾아진 항목집합의 수를 줄일 수 있다. 그 다음은 추정데이터 ((L_1^T)', (L_2^T)'), C_i(i=1, ..., Q)의 합집합($C_1 \cup C_2 ... \cup C_Q$)과 이질자 제거방법(trimming)을 이용해 최종샘플 S_0를 구한다. 끝으로 S_0에서 최소지지도와 최소신뢰도를 만족하는 이전 알고리즘보다 고품질의 연관규칙을 추출하게 된다.

3.2 다단계 샘플링 알고리즘

표 4. 다단계 샘플링 알고리즘

MultiFAST (D, N=|D|, n=|S_J|, J(단계의 수))
1. $\alpha = (N/n)^{\frac{1}{J}}$;
2. S_0 = D;
3. for (i=1; i>=J; i++) {
4. S_i = FAST(S_{i-1}, k, |S_{i-1}|, (|S_{i-1}|/α), *Dist*);
5. }
6. return S_J;

FAST 알고리즘과 위에서 제안된 복수 샘플링 알고리즘과 같은 trimming방식의 샘플링 알고리즘은 trimming과정에서 얻고자 하는 샘플의 상대적인 크기가 작을수록 전체 샘플링 시간은 더 걸린다는 것을 알 수 있다. 이점을 고려하여 J단계의 trimming 과정을 생각해 볼 수 있다. 다단계샘플링 알고리즘(MultiFAST)은 FAST 알고리즘을 J번 적용한다. 데이터베이스의 크기를 N, 원하는 최종샘플의 크기를 n, 전체 단계수를 J, 단축률을 α라고 하면 MultiFAST의 전체 알고리즘은 표 4와 같다. 부언하면 샘플데이터의 규모는 |D|/α^1 , |D|/α^2 , ..., |D|/α^J(=n) 와 같이 변화한다.

4. 결 론

FAST 알고리즘은 빈발 1-항목집합에 대해서는 이전 샘플링 기반의 연관규칙 추출알고리즘에 비해 정확도를 향상시켰지만 그 이상의 빈발항목집합의 정확도에 대한 고려는 하지 않았다. 데이터베이스내의 모든 빈발항목집합을 찾는 문제에 있어서 이러한 알고리즘은 유용하지 않을 수도 있다. 실제로 실험을 통해 FAST 알고리즘이 빈발 2-항목집합에서는 정확도가 떨어진다는 것을 확인하였다. 이러한 빈발 2-항목집합이상의 항목집합에 대한 정확도를 개선시키기 위해 본 논문에서는 복수 샘플링 알고리즘을 제안하였다. 현재 이 알고리즘은 구현이 진행되고 실험을 통해 그 유효성을 검증하고자 한다. 소개한 복수 샘플링 알고리즘을 통해 실행시간은 IFAST 알고리즘에 비해 늘어나지만 빈발 2-항목집합 이상의 항목집합에서 오류항목집합의 수를 줄임으로써 전체적인 정확도를 향상시킬 수 있는 샘플데이터를 얻을 것으로 기대하고 본격적인 응용에서는 병렬화를 통한 시간 단축을 꾀할 것 이다. 또한 다단계샘플링방식을 적용해 실행시간을 더욱 단축시킬 수 있을 것이다. 또한 복수 샘플링방식과 다단계샘플링방식의 알고리즘을 혼합하여 샘플링 기반의 연관규칙 마이닝 알고리즘의 정확도와 실행속도 모두를 개선시키기 위한 연구를 계획하고 있다.

참고문헌

[1] R. Agrawal, T. Imielinski, and A. Swami. "Mining Association Rules between Sets of Items in Large Databases", In ACM SIGMOD Intl. Conf. Management of Data, 1993.

[2] R. Agrawal and R. Srikant. "Fast algorithms for mining association rules". In Proc. VLDB Conf., 1994, pp.487-499.

[3] A. Savasere, E. Omiecinski, and S. Navathe, "An Efficient Algorithm for Mining Association Rules in Large Databases", In Proc. of the 24th VLDB Conference, 1995, pp.432-444.

[4] J. Han, J. Pei, and Y. Yin, "Mining Frequent Patterns without Candidate Generation", SIGMOD, 2000.

[5] M. El-hajj, O. Zaiane, "COFI-tree Mining: A New Approach to Pattern Growth with Reduced Candidacy Generation", FIMI, 2003.

[6] Toivonen, "Sampling Large Databases for Association Rules", In Proc. VLDB Conf., 1996.

[7] Bin Chen, Peter Haas, and Peter Scheuermann, "A new two-phase sampling based algorithm for discovering association rules", SIGKDD, 2002.

[8] 이문환, 김동승, "2단계 샘플링 방식의 정확성을 높인 연관성 추출 데이터 마이닝", 한국정보과학회 병렬처리시스템 학술대회 논문집, 16권 1호, pp.69-74, 2005.1월.