상관관계와 카이-제곱 분석에 기반한 긍정과 부정 연관 규칙 알고리즘

김나희*· 윤성대**

* 부경대학교 교육대학원 전산교육전공

** 부경대학교 전자컴퓨터정보통신공학부

Mining Positive and Negative Association Rules Algorithm based on Correlation and Chi-squared analysis

Na-hee Kim* · Sung-dae Youn**

* Dept. of Computer Education, Graduate school of PuKyung Nat'l University

** Div. of Electronic Computer And Telecommunication Engineering, Pukyung Nat'l University

E-mail : whitesky98@hanmail.net* sdyoun@pknu.ac.kr**

ABSTRACT

Recently, Mining negative association rules has received some attention and proved to be useful. Negative association rules are useful in market-basket analysis to identify products that conflict with each other or products that complement each other. Several algorithms have been proposed. However, there are some questions with those algorithms, for example, misleading rules will occur when the positive and negative rules are mined simultaneously. The chi-squared test that based on the mature theory and Correlation Coefficient can avoid the problem. In this paper, We proposed the algorithm PNCCR based on chi-squared test and correlation is proposed. The experiment results show that the misleading rules are pruned. It suggests that the algorithm is correct and efficient.

키워드

데이터마이닝, 연관 규칙, 부정 연관 규칙, 카이-제곱 분석, 상관관계, 알고리즘

Ⅰ. 서론

데이터 마이닝 기법 중에 많은 연구가 되고 있는 연관 규칙(association rule)은 하나의 거래나 사건에 포함되어 있는 둘 이상의 항목들을 파악해서 상호 관련성을 발견하는 것으로 대량의 트랜잭션에 존재하는 항목간의 암묵적인 관계를 찾아내는 작업을 말한다. 모든 전통적인 연관 규칙 마이닝 알고리즘들은 항목들 간의 긍정 연관 규칙(Positive Association rules; PARs)을 찾아내기 위해 개발되었다. 연관 규칙은 많은 학자들에 의해 연구되어 왔다.[1,2,3,4,5] 긍정 연관 규칙의 식은 X→Y 로 나타낸다. 여기에 중요한 추가 식으로 다음의 세 가지 식 X→¬Y, ¬X→Y, ¬X→¬Y가 있는데 이것은 부정 연관 규칙(Negative Association Rules; NARs)이라고 불린다. Brin 등에 의해 두 개의 빈발 항목 사이의 부정 관계에 대해 처음 언급된 이후로 최근 들어 부정 연관을 증명하는 것의 문제, 아이템의 부재를 조사하는 것이 연구되고 있으며[6], A. Savasere 등은 두 개의 빈발 항목 간에 강한 부정 상관관계를 제안하였다[7].

PARs와 NARs를 동시에 마이닝하게 되면 자기-모순 규칙(self-contradictory rules)이 같이 마이닝되는 문제점이 있다. 이들 자기-모순적인 규칙을 제거하기 위해서 여러 방법들이 제안되었다. 상관관계(Correlation)를 이용하여 잘못 유도된 규칙들을 제거한다[8]. 또한 interest measure를 이용하여 흥미 없는 규칙을 제거하였으며[9] 카이-제곱 분석을 하여 자기-모순적인 규칙을 제거하는 방법 등 여러 가지 방법이 제안되었다[10]. 모순 규칙의 문제는 NARs 연구에 있어서 매우 중요하다.

본 논문에서는 모순 규칙을 제거하고 도출된 규칙의 신뢰도 향상을 위해 상관관계(Correlation) 및 통계상의 이론에 기반을 둔 카이-제곱 분석(Chi-Squared analysis) 에 기반한 긍정, 부정 연관 규칙 알고리즘 PNCCR을 제안한다.

II. 관련연구

2.1 연관규칙

연관 규칙이란 동시에 발생하는 사건들을 규칙의 형태로 표현한 것으로 특정 사건이 발생하면 동시에 혹은 일정한 시간 간격 사이에 다른 사건이 일어나는 관계를 의미한다[11].

연관 규칙은 다음과 같이 정의된다. I={i_1, i_2, ..., i_n}으로 아이템들의 집합이다. 트랜잭션 T는 T⊆I인 항목들의 집합이고 D는 트랜잭션의 집합이다. 어떤 항목 집합을 X라고 했을 때, 트랜잭션 T가 필요충분조건으로 X⊆T를 만족하는 경우에만 트랜잭션 T가 항목 X를 포함한다고 한다. 각각의 X, Y가 X⊆I, Y⊆I이고, X∩Y=∅일 때, 연관규칙 X⇒Y의 형식으로 표현된다. 규칙 X⇒Y는 트랜잭션 집합 D에서 항목 X와 항목 Y를 동시에 포함하는 X∪Y 트랜잭션의 백분율이 s%일 때, 지지도(support) s를 가진다. 규칙 X⇒Y가 D에서 X를 포함하는 트랜잭션이 Y 또한 포함하고 있을 백분율이 c%라면 트랜잭션 집합 D에서 신뢰도(confidence) c를 갖는다고 한다.

2.2 부정 연관 규칙

지지도-신뢰도(support-confidence)의 기반에서 마이닝된 연관 규칙은 주어진 최소 지지도와 신뢰도를 만족한다. 그러나 실질적으로 결과 값은 잘못된 것일 수도 있다.

예를 들어, 장바구니에서 사과(A)와 바나나(B)의 트랜잭션을 분석한다고 가정하자. 10,000개의 트랜잭션에서 9,000개의 트랜잭션이 사과를 포함하고 있고, 2,500개의 트랜잭션이 바나나를 포함하고, 2,000개의 트랜잭션이 사과와 바나나를 둘 다 포함하고 있다. 표 1은 항목집합 A와 B의 우연성(contingency)을 나타낸다.

표 1. A와 B의 우연성(contingency)

	A	¬A	Σ
B	*2000($n_{A,B}$)*	*500($n_{,\neg AB}$)*	*2500($n_{.B}$)*
¬B	*7000($n_{\neg B,,A}$)*	*500($n_{\neg B,,\neg A}$)*	*7500($n_{.\neg B}$)*
Σ	*9000($n_{A.}$)*	*1000($n_{\neg A.}$)*	*10000(n)*

¬A는 항목 A의 부정으로써 항목 A의 부재를 의미한다. 최소 지지도를 0.2, 최소 신뢰도를 0.8이라고 가정했을 때 규칙 B⇒A는 유효한 연관 규칙이다. 그러나 두 항목의 상관관계 계수는 1보다 작은 0.89로써 사과와 바나나는 부정 연관이다.

부정 연관 규칙의 지지도는 다음 식을 통해 구한다.

$$supp(\neg X) = 1 - supp(X) \quad (1)$$

$$supp(X \cup \neg Y) = supp(X) - supp(X \cup Y) \quad (2)$$

$$supp(\neg X \cup Y) = supp(Y) - supp(X \cup Y) \quad (3)$$

$$\sup p(\neg X \cup \neg Y) = 1 - \sup p(X) - \sup p(Y) + \sup p(X \cup Y) \quad (4)$$

신뢰도 계산은 다음 식을 통해 할 수 있다.

$$conf(X \Rightarrow \neg Y) = \frac{\sup p(X) - \sup p(X \cup Y)}{\sup p(X)} \quad (5)$$

$$conf(\neg X \Rightarrow Y) = \frac{\sup p(X) - \sup p(X \cup Y)}{1 - \sup p(X)} \quad (6)$$

$$conf(\neg X \Rightarrow \neg Y) = \frac{1 - \sup p(X) - \sup p(Y) + \sup p(X \cup Y)}{1 - \sup p(X)} \quad (7)$$

2.3 상관관계

Positive와 Negative 연관 규칙을 동시에 마이닝하면 흥미롭지 않은 자기-모순적인 규칙(self-co-ntradictory rules)들이 많이 만들어진다. Correlat-ion은 이들 규칙을 제거하는데 사용된다. 사용하는 식은 다음과 같다.

$$corr_{X,Y} = \frac{\sup p(X \cup Y)}{\sup p(X) \sup p(Y)} \quad (8)$$

$corr_{X,Y}$ 의 값이 가지는 의미는 다음과 같다.

(1) $corr_{X,Y}$>1이면, 두 변수 X와 Y는 긍정 상관관계이다.

(2) $corr_{X,Y}$=1이면, 두 변수는 독립적이다.

(3) $corr_{X,Y}$<1이면, 두 변수는 부정 상관관계이다.

또한, $corr_{X,Y}$>1이면 $corr_{A,\neg B}$<1, $corr_{\neg A,B}$ <1, $corr_{\neg A,\ \neg B}$>1 의 관계가 있으며, 반대의 경우도 마찬가지이다.

2.4 카이-제곱 분석

카이-제곱 분석은 항목에 대한 가설 분석하기 위해 만들어진 일반적인 통계상의 방법이다. 이 분석방법은 두 개 혹은 더 많은 그룹의 샘플 비율들의 차이점을 분석할 수 있고 변수들 간의 연관성을 분석할 수 있다. 항목 집합들의 독립성은 카이-제곱의 수학적인 특성에 의해 확인되었다.

카이-제곱 분석의 주요 개념은 A와 B가 독립적이면 P(A∪B)=P(A)P(B)라는 것이다. 이 의미는 nAB=(nA.·nB)/n.이 되는데 다음 식을 통해 카이-제곱 값을 구할 수 있다.

$$x^2 = \frac{\left[n_{XY} - \frac{1}{n} n_Y * n_X\right]^2}{\frac{1}{n} n_Y * n_X} + \frac{\left[n_{\neg XY} - \frac{1}{n} n_Y * n_{\neg X}\right]^2}{\frac{1}{n} n_Y * n_{\neg Y}} + \frac{\left[n_{X\neg Y} - \frac{1}{n} n_{\neg Y} * n_X\right]^2}{\frac{1}{n} n_{\neg Y} * n_X} + \frac{\left[n_{\neg X\neg Y} - \frac{1}{n} n_{\neg Y} * n_{\neg X}\right]^2}{\frac{1}{n} n_{\neg Y} * n_{\neg X}} \quad (9)$$

가설 분석의 이론에 따라 아이템들 간의 연관을 판단한다. $X^2>X^2_{1-0.05}$이면 서로 관계가 있다.

III. 알고리즘 설계

3.1 알고리즘의 제안

긍정 연관 규칙과 부정 연관 규칙을 동시에 마이닝하여 규칙을 도출하는 경우에 자기-모순적인 규칙이 생성되는 문제가 있다. 모순 규칙의 문제는 NARs 연구에 있어서 중요한 문제이다. 상관관계 및 카이-제곱 분석은 변수간의 연관성을 찾아냄으로써 자기-모순적인 규칙을 제거하는 데에 사용된다. 이를 통해 잘못 유도된 규칙의 수를 줄이고 알고리즘의 효율성을 증명한다. 따라서 본 논문에서는 변수간의 상관관계와 카이-제곱 분석을 동시에 적용하여 흥미롭지 않은 규칙들을 제거하는 알고리즘 PNCCR을 제안한다. 알고리즘의 효율성을 증명하기 위해 상관관계를 이용한 알고리즘과 카이-제곱 분석을 사용한 알고리즘과 비교한다.

3.2 알고리즘

Algorithm : PNCCR(Positive and Negative Association Rules Generation based on Chi-Squared and Correlation)
Input TD, minsupp, minconf, and corr respectively Transactional Database, minimum support, minimum confidence, correlation.
Output AR: Positive and Negative Association Rules.
Method:
(1) *PAR*=∅; *NAR*=∅; /*positive and negative AR sets*/
(2) scan the database and find the set of frequent 1-itemsets (L_1)
(3) **for** (k=2, $L_{k-1}\neq 0$, k++){
(4) C_k= $L_{k-1} \bowtie L_{k-1}$
(5) **foreach** $i \in C_k$ {
(6) s=support(TD,i)
/*support of item i is computed*/
(7) **if** $s \geq minsupp$ **then**
(8) $L_k \leftarrow L_k \cup \{i\}$ /*item i is added to L_k*/
(9) **foreach** X,Y (i = X∪Y, X∩Y=∅) {
(10) *corr*=correlation(X,Y)
= sup(X∪Y) / (sup(X)sup(Y))
(11) **if** *corr*>1 **then**
(12) **if** $x^2 > x^2_{1,0.05}$ **then**
(13) **if** *sup(X->Y)-sup(X)sup(Y)>0* **then**
(14) **if** confidence(X→Y)≥ *minconf* **then**
(15) *PAR*← *PAR*∪{X→Y}
(16) **if** *confidence*(¬X→¬Y)≥*minconf* **then**
(17) *NAR* ←*NAR*∪{¬X→¬Y}
(18) **if** *corr*<1 **then**
(19) **if** $x^2 > x^2_{1-0.05}$ **then**
(20) **if** *sup(X->Y)-sup(X)sup(Y)<0* **then**
(21) **if** *confidence*(X→¬Y)≥*minconf* **then**
(22) *NAR*←*NAR*∪{X→¬Y}
(23) **if** *confidence*(¬X→Y)≥*minconf* **then**
(24) *NAR*←*NAR*∪{¬X→Y}
(25) }
(26) }
(27) }
(28) return PAR, NAR

상관관계를 비교하기 위해 항목 간의 상관관계를 계산한다(9-10행). 계산된 상관관계 값을 비교하여 1차적으로 상호-모순 규칙들을 정제한 긍정 연관규칙을 도출한다(11행). 다음에 카이-제곱 분석을 통해 상관관계로 정제하지 못한 자기-모순적인 규칙을 제거한다(12-13행). 이렇게 정제된 규칙들로 긍정 연관 규칙인 X⇒Y규칙을 찾아내고(14-15행) 상관관계가 가진 특성에 따라 부정 연관 규칙 ¬X⇒¬Y도 찾아낸다(16-17행). 다시 Correlation을 적용하여 남아있는 부정 연관 규칙을 도출하고(18행) 도출된 부정 연관 규칙에 카이-제곱 분석을 적용하여 규칙을 정제한다(19-20행). 정제된 규칙에서 부정 연관 규칙 X⇒¬Y와 규칙 ¬X⇒Y를 도출한다(21-24행).

IV. 실험결과

본 논문에서 제안한 알고리즘의 성능을 확인하기 위하여 인공의 데이터 집합에서 실험을 수행하였다. 데이터 집합은 50개의 항목, 100개의 트랜잭션을 유지한다. 상관관계 기반 알고리즘, 카이-제곱 분석 기반 알고리즘과 본 논문에서 제안한 상관관계와 카이-제곱 분석을 기반으로 한 알고리즘(PNCCR)을 비교하였다. minconf=8%라고 가정했을 때, 각각의 알고리즘은 연관규칙의 집합을 생성하기 위해 실행되어진다. 결과는 표 2에 나타난다.

표 2. 각 알고리즘이 도출해내는 규칙의 수

최소 신뢰도	규칙	규칙의 수		
		상관관계	카이-제곱 분석	PNCCR
10%	긍정연관	440	264	264
	부정연관	2136	1752	1639
20%	긍정연관	249	165	165
	부정연관	519	384	332

표 2를 보면 상관관계를 적용한 알고리즘보다 카이-제곱 분석을 적용한 알고리즘이 더 효율적이라 것을 확인할 수 있었고 본 논문에서 제안한 PNCRR 알고리즘은 카이-제곱 분석을 적용한 알고리즘보다 규칙의 수가 줄어드는 것을 확인할 수 있었다. 이것은 긍정 연관 규칙과 부정 연관 규칙을 동시에 마이닝하여 규칙을 도출하는 경우에 발생하는 자기-모순적인 규칙들이 제거되었다는 것을 의미한다.

Ⅴ. 결론

긍정 연관 규칙과 부정 연관 규칙을 동시에 마이닝하여 규칙을 찾아내는 경우에 자기-모순적인 규칙(self-contradictory rules)도 같이 도출되는 문제가 있다. 모순 규칙의 문제는 NARs 연구에 있어서 중요하다. 본 논문에서는 자기-모순적인 규칙을 제거하기 위하여 상관관계와 카이-제곱 분석에 기반한 PNCCR알고리즘을 제안하였다. 연관 규칙을 마이닝 할 때 잘못 유도되는 규칙들을 제거하기 위해 상관관계, 카이-제곱 분석 등의 매개 변수를 추가하는 방법들이 사용되어 왔다. 실험에서 각각의 방법들을 적용하여 알고리즘의 효율성을 비교하였고 카이-제곱 분석을 적용한 알고리즘이 규칙을 잘 정제한다는 것을 확인하였다. 또한 본 논문에서 제안한 PNCCR 알고리즘과 각 알고리즘들을 비교한 결과 제안한 알고리즘을 적용하였을 때 부정 연관 규칙에서 규칙의 수가 줄어드는 것을 확인하였다. 이것은 자기-모순적인 규칙들이 제거되었다는 것을 의미한다. 따라서 제안한 알고리즘이 더 효율적이라는 것을 증명하였다.

참고 문헌

[1] R. Agrawal, T. Imielinski, and A. Swami, “Mining Association Rules between Sets of Items in Large Databases”, ACM SIGMOD, pp. 207-216, 1993

[2] R. Agrawal and R. Srikant, “Fast Algorithms for Mining Association rules”, In Proceeding of the 20th Int’l Conf. on Very Large Databases(VLDB), Santiago, Chile, pp. 487-499, 1994.

[3] Jiawei Han, Yongjian Fu, “Mining multiple-level association rules in large databases”, Knowledge and Data Engineering, IEEE Transactions on Volume 11(5), pp. 798-805, 1999

[4] M. Klemettinen, H. Mannila, P. Ronkainen, H. Toivonen and A. I. Verkamo, “Finding interesting rules from large sets of discovered association rules”, In 3rd Int’l Conf on Information and Knowledge Management, pp. 401 - 407, 1994.

[5] R. Srikant, R. Agrawal. “Mining quantitative association rules in large relational tables”, ACM SIGMOD Record, 25(2), pp.1 - 12, 1996.

[6] S. Brin, R.Motwani, C. Silverstein, “Beyond Market: Generalizing Association Rules to correlations”, In Processing of the ACM SIGMOD Conference, pp. 265-276, 1997

[7] A. Savasere, E. Omiecinski, S. Navathe, “Mining for Strong Negative Associations in a Large Database of Customer Transactions”, In Proceedings of the 1998 International Conference on Data Engineering, pp.494-502, 1998

[8] X.J. Dong, S.J.Wang, H.T.Song et al, “Study ON Negative Association Rules”, Journal, Journal of Beijing Institute of Technology, China, pp. 978-981, 2004

[9] Shi-ju Shang, Xiang-jun Dong, Jie Li, Yuan-yuan Zhao, “Mining Positive and Negative Association Rules in Multi-database Based on Minimum Interestingness”, Intelligent Computation Technology and Automation (ICICTA), 2008 International Conference on Volume 1(20-22), pp.791-794, 2008

[10] Yuan-yuan Zhao, He Jiang, “Research of mining positive and negative weighted association rules based on Chi-squared analysis”, Information and Computing Science(ICIC), Second International Conference on Volume 1(21-22), pp. 344-347, May 2009

[11] Ke Kuo, Jie Wu, "A New method to Mine Vailed Association Rules", In Proceeding of the Second International Conference on Machine Learning and Cybernetics, Xi’an, 2-5, 2003