Proceedings of Korean Data And Information Science Society,
November 17-18, 2006, pp. 145-152

A Study for Statistical Criterion in Negative Association Rules Using Boolean Analyzer

Sang-Jin Shin[1], Keun-Woo Lee[2]

지도교수 : 박 희 창(창원대학교)

Abstract

Association rule mining searches for interesting relationships among items in a given database. Association rules are frequently used by retail stores to assist in marketing, advertising, floor placement, and inventory control. There are three primary quality measures for association rule, support and confidence and lift.

Association rule is an interesting rule among purchased items in transaction, but the negative association rule is an interesting rule that includes items which are not purchased. Boolean Analyzer is the method to produce the negative association rule using PIM. But, PIM is subjective. In this paper, we present statistical objective criterion in negative association rules using Boolean Analyzer.

keywords : association rule, Boolean Analyzer, PIM, negative association rule, statistical criterion,

1. 서론

데이터 마이닝 기법 중에 많은 연구가 되고 있는 연관성 규칙(association rule)은 하나의 거래나 사건에 포함되어 있는 둘이상의 품목들의 경향을 파악해서 상호 관련성을 발견하는 것으로 대용량 데이터베이스에 존재하는 항목간의 관련성을 찾아내는 작업을 말한다. 마케팅에서는 고객이 동시에 구매한 장바구니를 살펴봄으로써 거래되는 상품들의 관계를 발견 또는 분석한다는 의미에서 장바구니분석(market basket analysis)이라고 한다. 연관 규칙은 교차판매, 매장 진열, 카탈로그 디자인, 장바구니 분석 등에 사용된다. 각 항목간의 연관성을 반영하는 규칙으로 둘 또는 그 이상의 품목들 사이의 지지도(support), 신뢰도(confidence), 향상도(lift)를 바탕으로 관련성 여부를 측정한다. 연관 규칙은 탐색적이며, 비목적성 분석이며, 기존의 데이터를 특별한 변형 없이 계산이 용이하게 사용 가능하다는 장점을 가지고 있으며, 계산 과정이 길

1) Graduate Student, Department of Statistics, Changwon National University, Changwon, Gyeongnam, 641-773, Korea
E-mail : ssjpro@naver.com

2) Graduate Student, Department of Statistics, Changwon National University, Changwon, Gyeongnam, 641-773, Korea
E-mail : woolee22@hotmail.com

고, 반복된 계산이 많으며, 적절한 품목의 결정이 어렵고, 각 품목의 단위에 따른 표준화가 어렵다는 단점을 아울러 가지고 있다. 연관 규칙은 이러한 단점에도 불구하고 두 품목간의 관계를 명확히 수치화함으로써 두 개 이상의 품목간의 관련성을 나타내주기 때문에 현업에서 많이 활용되고 있다. 연관성 규칙은 Agrawal 등(1993)에 의해 처음 소개된 이후, 많은 학자들에 의해 연구되고 있다(Agrawal 등(1994), Cheung 등(1996), Park 등(1995), Sergey 등(1997), Toivonen(1996), Saygin 등(2002) 등).

일반적으로 연관성 규칙은 상품의 동시 구매에 관한 연관 규칙에 관심을 갖는다. 그러나 연관 규칙보다 음의 연관 규칙(negative association rule)에 관심을 갖는 경우가 종종 발생할 수 있다. 기존의 연관 규칙이 $A \rightarrow B$ 규칙과 같이 A상품을 사는 사람들은 B상품을 사는 규칙을 의미한다면, 음의 연관 규칙은 $\sim A \rightarrow \sim B$ 규칙과 같이 A상품을 사지 않는 사람들은 B상품을 사지 않는 규칙을 의미한다. 이외에도 음의 연관 규칙에는 $\sim A \rightarrow B$, $A \rightarrow \sim B$ 형태의 규칙이 있으며 어느 한쪽에 'not'이 들어간 규칙을 의미한다. 음의 연관 규칙은 연관규칙과 같이 흔하게 나타나는 규칙은 아니지만 지지도와 신뢰도가 음의 연관 규칙에서 훨씬 높게 나타난다면 오히려 음의 연관규칙에서 찾아낸 규칙이 훨씬 더 가치 있다고 볼 수 있다(이종인(2003)).

음의 연관 규칙을 찾는 연관성 규칙 알고리즘으로는 Agrawal 등(1993)의 Apriori 알고리즘과 Yung(2002)의 Taxonomy 알고리즘 등이 있다. 그러나 음의 연관 규칙을 찾기 위한 기존의 알고리즘들은 규칙의 일부분만 찾거나, 규칙을 찾는데 매우 복잡한 과정을 거치므로 비용과 시간이 많이 드는 문제점이 있다.

Orchard(1975)는 PIM(probabilistic interestingness measure)을 이용한 Boolean Analyzer 알고리즘을 제시하였고 Imberman 등(2002)이 머리 외상 데이터에 이를 적용하여 연관성 규칙을 생성하였다. Boolean Analyzer 알고리즘은 확률을 이용하여 각 아이템들 사이의 의존성을 계산하여 서로 얼마나 연관이 있는지를 찾아내는 방법이다. Boolean Analyzer 알고리즘은 기존의 방법보다 간단한 과정으로 음의 연관 규칙을 생성할 수 있다. Boolean Analyzer에서는 사건의 확률에 근거해서 연관성의 정도를 나타내는 PIM을 만들어 내고 PIM을 이용하여 음의 연관 규칙을 생성한다. 여기서 PIM 값이 0이 아닐 때 연관 관계가 존재한다고 할 수 있다. 그러나 PIM 값이 어느 정도 일 때 '관련성이 있다'라고 판단하는 객관적 기준이 없다. 이에 본 논문에서는 보다 객관적인 PIM의 연관 기준값을 제시하는 방법에 대하여 연구하고자 한다. 본 논문의 2절에서는 Boolean Analyzer를 이용한 음의 연관 규칙에서의 통계적 결정기준에 대하여 기술하고 3절에서는 모의실험 결과에 대하여 기술하며, 4절에서 결론을 맺는다.

2. Boolean Analyzer를 이용한 음의 연관규칙에서의 통계적 결정기준

연관 규칙은 항목 집합으로 표현된 트랜잭션에서 각 항목간의 연관성을 반영하는 규칙으로서, 기본적으로 미리 결정된 최소 지지도 이상의 트랜잭션 지지도를 가지는 항목 집합들의 모든 집합들인 빈발 항목 집합들을 찾아내어 연관 규칙을 생성한다. 두 항목 간의 연관 규칙은 $A \rightarrow B$로 표현되며 '항목 A가 발생하면 항목 B도 발생한다.'고 해석할 수 있다.

음의 연관 규칙은 기존의 연관 규칙에서 'not'의 개념이 들어간 것이라고 볼 수 있

다. 음의 연관 규칙을 기호로 나타내면 $\sim A \rightarrow \sim B$, $\sim A \rightarrow B$, $A \rightarrow \sim B$의 형태로 표현할 수 있으며, '~'는 'not'의 의미이다. 음의 연관 규칙은 연관 규칙과 같이 흔하게 나타나는 규칙이 아니다. 그러나 만일 생성된 규칙 중 지지도와 신뢰도가 음의 연관 규칙 쪽에서 훨씬 높게 나타난다면 오히려 음의 연관 규칙에서 찾아낸 규칙이 훨씬 더 가치가 있을 것이다. 음의 연관 규칙은 연관 규칙의 활용과 마찬가지로 고객의 구매 경향 및 마케팅 정책을 제시할 수 있고 교차판매, 매장 진열, 카탈로그 디자인 등의 타겟 마케팅(target marketing)등에 활용 가능하다.

음의 연관 규칙을 찾는 연관성 규칙 알고리즘으로는 역 데이터베이스를 이용한 Apriori 알고리즘과 item들의 grouping을 이용하는 Taxonomy 알고리즘 등이 있으나 몇 가지 문제점이 있다. Apriori 알고리즘은 구매하지 않은 데이터를 관찰하기 위해 기존의 데이터베이스를 전부 역 데이터베이스로 변환하는 작업이 필요하다. 변환된 역 데이터베이스는 전체 item의 숫자에 비례하여 증가하므로 연관 규칙을 찾기 위해서 사용한 데이터의 수와 비교해 본다면 역 데이터베이스를 이용하여 음의 연관 규칙을 찾는 것은 비용이 상당히 많이 드는 단점이 있다. Taxonomy 알고리즘은 불필요한 규칙을 찾는데 비용을 들이지 않고 유용하고 정확한 규칙을 찾아낼 수 있다는 장점이 있다. 그러나 음의 연관규칙을 찾으려는 과정이 복잡하다. 우선 Apriori 알고리즘으로 연관 규칙을 찾아야 하고 찾은 결과를 이용하여 음의 연관규칙을 찾은 뒤, 중복된 규칙을 제거해야하므로 연관 규칙만을 찾는 것보다 과정이 더 복잡하다.

Boolean Analyzer 알고리즘은 기존의 알고리즘에 비하여 보다 간단한 계산과정으로 음의 연관 규칙을 발견할 수 있다. Boolean Analyzer 알고리즘은 사건의 확률에 근거해서 연관성의 정도를 나타내는 PIM을 만들어 내고 PIM을 이용하여 연관성 규칙을 생성한다. Boolean Analyzer 알고리즘은 PIM의 값의 크기로 연관의 정도를 나타낸다. <표 1>과 같이 X와 Y를 하나의 사건이라고 간주하고 'X'와 'Y"는 일어나지 않는 사건의 경우라고 하자.

<표 1> 2X2 테이블

	Y	Y′
X	a	b
X′	c	d

<표 1>에서와 같이 테이블의 각 값을 a, b, c, d로 놓으면 PIM은 식(2.1)과 같다.

$$\Pi M = ad - bc \tag{2.1}$$

PIM은 연관 정도를 수치적으로 표현 가능하게 하는 척도로서 연관 정도의 순위까지도 알 수 있다. <표 1>에서 a, b, c, d가 서로 독립이면 PIM이 0이 되어 아무런 연관 관계가 없는 것으로 판단한다. a, b, c, d가 독립이 아닐 경우 PIM이 0이 아닌 값이 나타나므로 이를 통해 연관 정도를 확인 할 수 있다. PIM은 사건 X와 Y의 의존의 정도를 나타내는 측정치이다. Imberman 등(2002)에 의하면 PIM값이 양수 값이면 강한 의존성 관계를 나타낸다. 즉, 두 사건 간에는 강한 연관성이 존재한다는 의미이다. PIM 값이 0이나 0에 가까운 값이면 두 사건은 서로 독립이므로 아무런 연관 관계가

없음을 나타낸다. PIM 값이 음수이면 음의 연관 관계를 나타낸다고 할 수 있다. 그러나 PIM 값이 어느 정도 일 때 '관련성이 있다'라고 판단하는 객관적 기준이 없어 연구자가 임의로 그 기준을 정하고 있다. 이에 본 절에서는 보다 객관적인 PIM의 연관 기준값을 제시하는 방법을 기술하고자 한다.

$r \times c$ 분할표에서 i행 j열째의 실측도수를 f_{ij}, 기대도수를 t_{ij}라 하면 카이제곱 검정 통계량 값은 식 (2.2)와 같이 표현된다.

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(f_{ij} - t_{ij})^2}{t_{ij}} \tag{2.2}$$

여기서 자유도는 $(r-1)\times(c-1)$이다.

식 (2.2)를 $r \times c$ 분할표가 아닌 <표 1>을 이용하여 PIM을 식으로 변형하여 나타내면 식(2.3)이 된다.

$$IIM = \frac{n}{2} + \sqrt{\frac{(a+b)(a+c)(b+d)(c+d)\chi^2}{(a+b+c+d)}} \tag{2.3}$$

여기서 χ^2의 자유도는 1이다.

식 (2.3)에서 χ^2 값이 통계적으로 유의한지 아닌지를 판단할 수 있으므로 객관적인 PIM의 연관 기준값을 제시한다고 할 수 있다.

3. 모의실험

본 장에서는 다음과 같이 모의실험을 실시하였다. 실험은 각각 n, y_1, x_1을 모두 고정한 후 a의 범위를 구하고, a의 변화에 따라 PIM 값과 카이제곱 검정 결과의 유의성을 규명하여, PIM의 객관적 연관 기준값을 알아보고자 한다. 모의실험 데이터의 구조는 <표 2>과 같다.

<표 2> 모의실험 데이터

		Y		합
		O	X	
X	O	a	30- a	30
	X	50- a	a+25	70
합		50	50	100

$n=100$, $x_1=30$, $y_1=50$를 적용하였을 경우, a가 취할 수 있는 정수 값의 범위

는 다음과 같다.

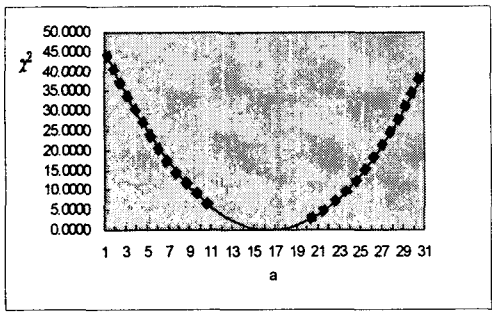
$$0 \leqq a \leqq 30$$

모의실험 결과 <표 3>과 같이 a값에 따른 PIM 값과 카이제곱 통계량 값을 확인할 수 있다($\alpha = 0.05$, $\chi^2(1) = 3.84146$). <표 3>에서는 χ^2값이 통계적으로 유의한 부분에 대하여 음영으로 처리하였다.

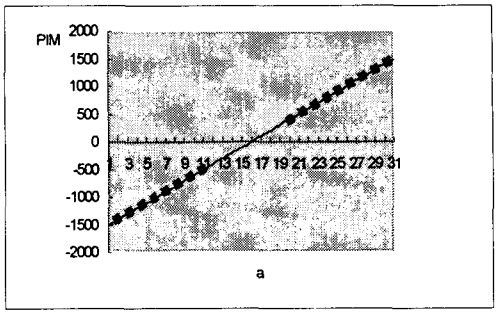
<표 3> a에 따른 PIM과 카이제곱

a	b	c	d	PIM	χ^2	Support	Confidence
1	29	49	21	-1400	40.0476	0.01	0.033
2	28	48	22	-1300	34.7143	0.02	0.067
3	27	47	23	-1200	29.7619	0.03	0.100
4	26	46	24	-1100	25.1905	0.04	0.133
5	25	45	25	-1000	21.0000	0.05	0.167
6	24	44	26	-900	17.1905	0.06	0.200
7	23	43	27	-800	13.7619	0.07	0.233
8	22	42	28	-700	10.7143	0.08	0.267
9	21	41	29	-600	8.0476	0.09	0.300
10	20	40	30	-500	5.7619	0.10	0.333
11	19	39	31	-400	3.8571	0.11	0.367
12	18	38	32	-300	2.3333	0.12	0.400
13	17	37	33	-200	1.1905	0.13	0.433
14	16	36	34	-100	0.4286	0.14	0.467
15	15	35	35	0	0.0476	0.15	0.500
16	14	34	36	100	0.0476	0.16	0.533
17	13	33	37	200	0.4286	0.17	0.567
18	12	32	38	300	1.1905	0.18	0.600
19	11	31	39	400	2.3333	0.19	0.633
20	10	30	40	500	3.8571	0.20	0.667
21	9	29	41	600	5.7619	0.21	0.700
22	8	28	42	700	8.0476	0.22	0.733
23	7	27	43	800	10.7143	0.23	0.767
24	6	26	44	900	13.7619	0.24	0.800
25	5	25	45	1000	17.1905	0.25	0.833
26	4	24	46	1100	21.0000	0.26	0.867
27	3	23	47	1200	25.1905	0.27	0.900
28	2	22	48	1300	29.7619	0.28	0.933
29	1	21	49	1400	34.7143	0.29	0.967
30	0	20	50	1500	40.0476	0.30	1.000

<표 3>에서 보는 바와 같이 a가 증가할수록 PIM의 값이 증가함을 알 수 있다. a의 값이 11 이하 일 때 PIM의 값이 음수이면서 카이제곱 통계량이 유의함을 알 수 있고, a가 20 이상 일 때 PIM 값이 양수이면서 카이제곱 통계량이 유의함을 알 수 있다. 이를 그림으로 나타내면 <그림 1> 및 <그림 2>와 같다. <그림 1> 및 <그림 2>에서는 통계적으로 유의한 부분에 대하여 굵은 점선으로 표시하였다.



<그림 1> 카이제곱 통계량 변화량



<그림 2> PIM 변화량

<그림 1>과 <그림 2>는 a에 따른 카이제곱 통계량 값과 PIM 값의 변화를 보여주고 있다. 그림에서 보는 바와 같이 a가 11 이하와 20 이상 일 때 카이제곱 통계량 값이 유의함을 알 수 있으며, a가 11 이하일 때 음의 의존성을 나타내고 있음을 알 수 있다. 이에 음의 연관 규칙을 생성하기 위한 PIM의 객관적 연관 기준값은 -400임을 알 수 있다.

4. 결론

연관성 규칙은 둘이상의 품목들의 경향을 파악해서 상호 관련성을 발견하고 분석하는 방법이다. 일반적으로 연관성 규칙은 상품의 동시 구매에 관한 연관 규칙에 관심을 갖는다. 그러나 연관 규칙보다 음의 연관 규칙에 관심을 갖는 경우가 발생할 수 있다. 음의 연관 규칙 알고리즘으로는 Apriori 알고리즘과 Taxonomy 알고리즘 등이 있다. 그러나 음의 연관 규칙을 찾기 위한 기존의 알고리즘들은 규칙의 일부분만 찾거나 매우 복잡한 과정을 거치므로 비용과 시간이 많이 드는 문제점이 있다.

Boolean Analyzer는 기존의 방법보다 간단한 과정으로 음의 연관 규칙을 생성할 수 있다. Boolean Analyzer 알고리즘은 확률에 근거해서 의존성의 정도를 나타내는 PIM을 만들지만 PIM의 객관적인 기준이 없다. 이에 본 논문은 PIM의 객관적인 연관 기준값을 제시하는 방법을 연구하였고 모의실험을 통하여 이를 확인하였다. 향후 과제로 연관성 규칙에 사용되는 여러 가지 척도에 대하여 본 논문에서 제안한 객관적 기준을 제시하는 방법을 적용하고자 한다.

참고 문헌

1. 이종인, 박상호, 강윤희, 박선, 이주홍(2003), Boolean Analyzer를 이용한 역 연관규칙의 발견, 2003년도 한국정보과학회 가을 학술발표논문집, 제 30권, 2호, pp.187-189.
2. Agrawal R., Imielinski R., Swami A.(1993), Mining association rules between sets of items in large databases, *In Proc. of the ACM SIGMOD Conference on Management of Data*, Washington, D.C.
3. Agrawal R., Srikant R.(1994), Fast algorithms for mining association rules, *In Proc. of the 20th VLDB Conference*, Santiago, Chile.
4. Cheung D.W., Han J., Ng V., Fu A.W., Fu Y.(1996), A Fast distributior algorithm for mining association rules, *Int's Conf. on Parallel and Distributes Information System*, Miami Beach, Florida.
5. Orchard R.A.(1975), On the Determination of Relationships Betweer Computer System State Variables.
6. Park J.S., Chen M.S., and Philip S.Y.(1995), An effective hash-based algorithms for mining association rules, *In Proc. of ACM SIGMOD Conference on Management of Data.*, Washington, D.C.
7. Saygin Y., Vassilios S.V., Clifton C.(2002), Using Unknowns to Prevent Discovery of Association Rules, *2002 Conf. on Research Issues in Data Engineering.*
8. Sergey B., Rajeev M., Jeffrey D.U., Shalom T.(1997), Dynamic itemset counting and implication rules for market data, *In Proc. of ACM SIGMOD Conference on Management of Data.* Washington, D.C.
9. Imberman S., Domanski B., Thompson H.(2001), Boolean Analyer - Ar Algorithm That Uses A Probabilistic Interestingness Measure to find Dependency/Association Rules In A Head Trauma Data, *A Workshop at the Tenth World Congress on Health and Medical Informatics.*
10. Toivonen H.(1996), Sampling Large Database for Association Rules, *Proc. of the 22nd VLDB Conference Mumbai(Bombay)*, India.
11. Yuan X.(2002), Mining Negative Association Rules, *International Symposium on Computer and Communications*, Vol. 2002, pp.623-628.