Proceedings of Korean Data And Information Science Society,
November 17-18, 2006, pp. 197-204

Exploiting a statistical threshold for efficiently identifying correlated pairs

Myoung-Ju Kim[1], Hee-Chang Park[2]

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. If there is many item in the association rule, much time is required. Xiong(2004) studies new method which is to compute the support of upper. They used support of upper to the θ. But θ is subjective. In this paper, we present statistical objective criterion for efficiently identifying correlated pairs.

keywords : association rule, correlation coefficient, statistical threshold, support, upper bound

1. 서론

데이터마이닝은 방대한 양의 데이터 속에서 쉽게 드러나지 않는 유용한 정보를 찾아내는 과정이다. 데이터마이닝은 대용량(massive)의 관측 가능한 데이터를 기반으로 숨겨진 지식, 기대하지 못했던 패턴, 새로운 법칙과 관계를 발견하고 이를 바탕으로 의사결정 등을 위한 정보로 활용하는 것이다. 데이터마이닝 기법으로는 군집분석(cluster analysis), 연결 분석(link analysis), 판별 분석(discrimination analysis), 연관성규칙(association rule), 의사결정나무(decision tree) 기법, 신경망모형(neural network) 등의 분석 기법이 있다.

본 논문에서 사용한 연관성 규칙은 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. 연관성규칙은 탐색적이며, 비목적성 분석이며, 기존의 데이터를 특별한 변형 없이 계산이 용이하게 사용 가능하다는 장점을 가지고 있으며, 계산 과정이 길고, 반복된 계산이 많으며, 적절한 품목의

1) Graduate Student, Department of Statistics, Changwon National University, Changwon, Gyeongnam, 641-773, Korea
E-mail : koongmul@hanmail.net
2) Corresponding author : Professor, Department of Statistics, Changwon National University, Changwon, Gyeongnam, 641-773, Korea
E-mail : hcpark@changwon.ac.kr

결정이 어렵고, 각 품목의 단위에 따른 표준화가 어렵다는 단점을 아울러 가지고 있다. 연관성규칙은 이러한 단점에도 불구하고 두 품목간의 관계를 명확히 수치화(지지도, 신뢰도, 향상도)함으로써 두 개 이상의 품목간의 관련성을 표시하여 주기 때문에 현업에서 많이 활용되고 있다.

연관성규칙은 Agrawal 등(1993)에 의해 처음 소개된 이후, Agrawal 등(1994)은 후보 항목 집합을 구성하고, 발생 빈도수를 계산하고 난 후에 사용자가 정의한 최소 지지도를 기초로 빈발 항목 집합들을 결정하는 Apriori, AprioriTid 알고리즘을 제안하였다. Park 등(1995)은 partitioning 알고리즘을 제안하였으며, Toivonen(1996)은 sampling 알고리즘을 제안하였다. 또한 Cheung 등(1996)은 FUP(fast update) 알고리즘에 대한 연구를 하였고, Sergey 등(1997)은 DIC(dynamic itemset counting)등의 발전된 연관성 규칙 알고리즘들이 연구되었다.

연관성 규칙 생성에서 항목의 수가 많을 경우, 각 항목 간 처리의 수가 크게 늘어 규칙 생성에 있어 많은 시간이 소요될 것이다. 이에 Xiong(2004)은 피어슨의 상관계수를 이용하여 상한의 지지도를 계산한 후, 높은 양의 상관을 가진 모든 항목 쌍들에 대해서만 연관 규칙을 생성하는 방법을 제시하여 효율적인 계산이 가능함을 보여주고 있다. 여기서 Xiong(2004)은 상관계수를 이용하여 상관계수에 대한 상한의 지지도를 계산하고 연구자가 임의로 정한 상한의 지지도 θ를 통하여 연관성 규칙을 생성하고 있다. 그러나 상한의 지지도 θ를 연구자가 임의로 지정하고 있어 θ에 대한 객관적인 기준이 없다. 이에 본 논문에서는 강한 연관성 규칙을 찾기 위해 사용되는 범위 θ에 대한 객관적인 기준을 제시하는 방법을 연구하고자 한다.

본 논문의 2절에서는 강한 연관성 확인을 위한 객관적 기준에 대하여 기술하고 3절에서는 모의실험에 대하여 기술하며, 4절에서 결론을 맺는다.

2. 강한 연관성 확인을 위한 객관적 기준

Xiong(2004)은 효율적인 규칙 생성을 위하여 각 항목 간의 모든 조합에 대한 관련성을 계산하지 않고 피어슨의 상관계수를 이용하여 상한의 지지도를 계산한 후, 높은 양의 상관을 가진 모든 항목 쌍들에 대해서만 관련성을 계산하고 있다. Xiong(2004)이 제시하고 있는 방법은 다음과 같다. <표 1>과 같이 2×2 테이블이 있다고 하자.

<표 1> 항목 A와 항목 B에 대한 2×2 테이블

		B		열의 합
		0	1	
A	0	$P_{(00)}$	$P_{(01)}$	$P_{(0+)}$
	1	$P_{(10)}$	$P_{(11)}$	$P_{(1+)}$
행의 합		$P_{(+0)}$	$P_{(+1)}$	N

여기서 $P_{(ij)}$ $(i=0,1;\ j=0,1)$는 i번째 행과 j번째 열의 표본의 크기이다. 또한 $P_{(i+)}$와 $P_{(+j)}$는 각각 i번째 항의 전체 표본의 크기와 j번째 열의 전체 표본의 크기이다.

<표 1>에 대한 상관계수를 계산하면 식(2.1)과 같다.

$$\Phi = \frac{P_{(00)}P_{(11)} - P_{(01)}P_{(10)}}{\sqrt{P_{(0+)}P_{(1+)}P_{(+0)}P_{(+1)}}} \tag{2.1}$$

식(2.1)을 변형하면 식(2.2)와 같이 표현할 수 있다.

$$\Phi = \frac{\frac{P_{(11)}}{N} - \frac{P_{(1+)}}{N}\frac{P_{(+1)}}{N}}{\sqrt{\frac{P_{(0+)}}{N}\frac{P_{(1+)}}{N}\frac{P_{(+0)}}{N}\frac{P_{(+1)}}{N}}} \tag{2.2}$$

식(2.2)에서 나타난 각 항을 지지도(support)로 나타내면 supp(A) = $P_{(1+)}/N$, supp(B) = $P_{(+1)}/N$ 이고 supp(A,B) = $P_{(11)}/N$이므로 상관계수를 지지도로 나타내면 식(2.3)과 같다.

$$\Phi = \frac{supp(A,B) - supp(A)\,supp(B)}{\sqrt{supp(A)\,supp(B)\,(1-supp(A))\,(1-supp(B))}} \tag{2.3}$$

식(2.3)에서 supp(A) ≥ supp(B)이라고 가정하면 항목쌍 {A,B}에 대한 상관계수의 상한 upper($\Phi_{\{A,B\}}$)은 supp(A,B)=supp(B)이고 식(2.4)와 같이 표현된다.

$$\text{upper}(\Phi_{\{A,B\}}) = \sqrt{\frac{supp(B)}{supp(A)}}\sqrt{\frac{1-supp(A)}{1-supp(B)}} \tag{2.4}$$

Xiong(2004)은 식(2.4)의 상관계수의 상한 upper($\Phi_{\{A,B\}}$)를 θ로 지정하고 지정된 θ보다 작은 값을 가지는 경우 연관성 계산을 하지 않아 계산의 양을 절약하고 있다. 그러나 상관계수의 상한인 θ를 연구자가 임의로 지정하고 있어 θ를 잘못 지정할 경우 불필요한 규칙을 계산할 수 있으며 중요한 규칙을 발견하지 못할 수도 있다. 이에 본 절에서는 상관계수의 검정통계량인 T값을 이용하여 θ의 객관적인 기준을 제시하

고자 한다. θ의 객관적인 기준을 제시하는 방법은 다음과 같다.

상관계수의 검정통계량 값은 식(2.5)와 같다.

$$T = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \geq t_{\frac{\alpha}{2}}(n-2) \tag{2.5}$$

식(2.5)를 상관계수 r로 표현하면 식(2.6)과 같다.

$$r \geq \frac{t_{\frac{\alpha}{2}}(n-2)}{\sqrt{(n-2) + (t_{\frac{\alpha}{2}}(n-2))^2}} \tag{2.6}$$

상관계수의 상한인 θ를 지정하면 식(2.7)과 같다.

$$\sqrt{\frac{supp(B)}{supp(A)}}\sqrt{\frac{1-supp(A)}{1-supp(B)}} \leq \Theta \tag{2.7}$$

식(2.7)을 다시 정리하면 식(2.8)과 같이 표현된다.

$$supp(B) > \frac{supp(A)}{\Theta^2 + (1-\Theta^2)\,supp(A)} \tag{2.8}$$

식(2.8)에 식(2.6)을 대입하면 식(2.9)와 같이 표현된다.

$$supp(B) > \frac{supp(A)}{\left\{\frac{t_{\frac{\alpha}{2}}(n-2)}{\sqrt{(n-2) + (t_{\frac{\alpha}{2}}(n-2))^2}}\right\}^2 + \left(1 - \left\{\frac{t_{\frac{\alpha}{2}}(n-2)}{\sqrt{(n-2) + (t_{\frac{\alpha}{2}}(n-2))^2}}\right\}^2\right)supp(A)} \tag{2.9}$$

$\Theta = t_{\frac{\alpha}{2}}(n-2)/\sqrt{(n-2) + (t_{\frac{\alpha}{2}}(n-2))^2}$이므로 이를 n에 관한 식으로 변형하면 식(2.10)과 같다.

$$n = \frac{-\{t_{\frac{\alpha}{2}}(n-2)\}^2}{\Theta - \{t_{\frac{\alpha}{2}}(n-2)\}^2} + 2 \tag{2.10}$$

식(2.10)과 같이 표본의 크기 n이 지정되면 상관계수의 상한인 θ의 객관적인 기준

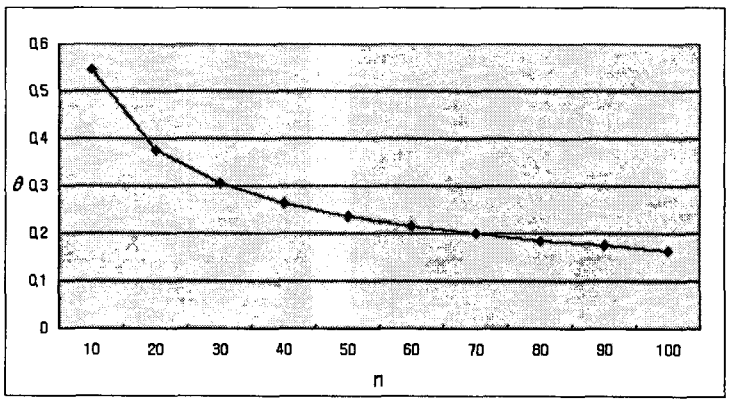
이 제시된다. 강한 연관성 규칙을 확인하기 위하여 상관계수의 상한인 θ 값을 결정할 때 식(2.10)에서 구해진 θ값을 지정하는 것이 바람직하다.

3. 모의실험

본 장에서는 다음과 같이 모의실험을 실시하였다. 이 모의실험은 n을 변화시켰을 때 각각의 객관적 Θ의 값(유의수준 : 0.05)을 구하고자 한다. n을 10에서 100으로 변화하였을 때의 모의실험 결과는 <표 2> 및 <그림 1>과 같다.

<표 2> 모의실험 1의 결과

n	$t_{\frac{\alpha}{2}(n-2)}$	Θ
10	2.306	0.544
20	2.101	0.377
30	2.048	0.306
40	2.021	0.264
50	2.009	0.236
60	2.000	0.214
70	2.000	0.199
80	1.990	0.186
90	1.990	0.175
100	1.984	0.165



<그림 1> 모의실험 1의 Θ 변화량

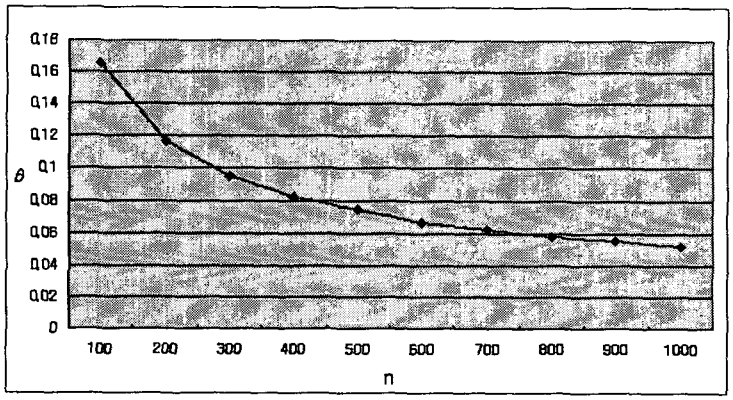
<표 1> 및 <그림 1>에서 보는 바와 같이 n값에 따른 Θ의 객관적 기준이 제시되

고 있다. n이 10일 때 Θ의 객관적 기준은 0.544가 되고 n이 100일 때 Θ의 객관적 기준은 0.165가 된다.

다음으로 n을 100에서 1000으로 변화하였을 때의 모의실험 결과는 <표 3> 및 <그림 2>와 같다. <표 3> 및 <그림 2>에서 보는 바와 같이 n이 200일 때 Θ의 객관적 기준은 0.117이 되고 n이 1000일 때 Θ의 객관적 기준은 0.052가 된다. 모의실험 결과 n이 커지면 커질수록 Θ의 값은 줄어드는 것을 알 수 있었다.

<표 3> 모의실험 2의 결과

n	$t_{\frac{\alpha}{2}(n-2)}$	Θ
100	1.984	0.165
200	1.984	0.117
300	1.984	0.095
400	1.984	0.082
500	1.984	0.074
600	1.962	0.067
700	1.962	0.062
800	1.962	0.058
900	1.962	0.055
1000	1.962	0.052



<그림 2> 모의실험 2의 Θ 변화량

4. 결론

연관성 규칙은 하나의 거래나 사건에 포함되어 있는 둘이상의 품목들의 경향을 파

악해서 상호 관련성을 발견하는 것으로 대용량 데이터베이스에 존재하는 항목간의 관련성을 찾아내는 방법이다. 그러나 연관성 규칙 생성에서 항목의 수가 많을 경우, 각 항목 간 처리의 수가 크게 늘어 규칙 생성에 있어 많은 시간이 소요된다. 이에 Xiong(2004)은 피어슨의 상관계수를 이용하여 상한의 지지도를 계산한 후, 높은 양의 상관을 가진 모든 항목 쌍들에 대해서만 연관 규칙을 생성하는 방법을 제시하여 효율적인 계산이 가능함을 보여주고 있다. 여기서 상한의 지지도 θ를 연구자가 임의로 지정하고 있어 θ에 대한 객관적인 기준이 없다. 이에 본 논문에서는 강한 연관성 규칙을 찾기 위해 사용되는 범위 θ에 대한 객관적인 기준을 제시하는 방법을 연구하였고 모의실험을 통하여 이를 확인하였다. 향후 연구과제로 실제 자료에 본 연구 방법을 적용하여 그 결과를 확인하고 수행 시간을 비교하는 연구를 할 필요성이 있다.

참고 문헌

1. 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.
2. Agrawal R., Srikant R.(1994), Fast algorithms for mining association rules, *In Proc. of the 20th VLDB Conference*, Santiago, Chile.
3. Cheung D.W., Han J., Ng V., Fu A.W., Fu Y.(1996), A Fast distribution algorithm for mining association rules, *Int's Conf. on Parallel and Distributes Information System*, Miami Beach, Florida.
4. Xiong H., Shekhar S., Tan P., Kumar V.(2004), Exploiting A Support-based Upper Bound of Pearson's Correlation Coefficient for Efficiently Identifying Strongly Correlated Pairs, *In Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.*
5. 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.
6. Sergey B., Rajeev M., Jeffrey D.U., Shalom T.(1997), Dynamic itemset counting and implication rules for market data, *In Proceedings of ACM SIGMOD Conference on Management of Data.* Washington, D.C.
7. Toivonen H.(1996), Sampling Large Database for Association Rules, *Proceedings of the 22nd VLDB Conference Mumbai(Bombay)*, India.