전략적 중요도를 고려한 연관규칙 탐사

최덕원, 신진규
*성균관대학교 시스템경영공학과

e-mail: dougch01@paran.com, sjg0311@paran.com,

Association Rule Mining Considering Strategic Importance

Doug Won Choi, Jin-Gyu Shin

Department of Systems Management Engineering, SungKyunKwan University

Abstract

A new association rule mining algorithm, which reflects the strategic importance of associative relationships between items, was developed and presented in this paper. This algorithm exploits the basic framework of Apriori procedures and TSAA(transitive support association Apriori) procedure developed by Hyun and Choi in evaluating non-frequent itemsets. The algorithm considers the strategic importance(weight) of feature variables in the association rule mining process. Sample feature variables of strategic importance include: profitability, marketing value, customer satisfaction, and frequency.

A database with 730 transaction data set of a large scale discount store was used to compare and verify the performance of the presented algorithm against the existing Apriori and TSAA algorithms. The result clearly indicated that the new algorithm produced substantially different association itemsets according to the weights assigned to the strategic feature variables.

01. 서론

본 논문에서는 가중치의 개념을 이용하여 기존의 연관규칙 방법론(Apriori)이 갖고 있는 한계를 극복하는 새로운 방법론의 개발에 초점을 두었다. 단순히 데이터의 빈발도만을 이용하여 연관규칙을 생성하게 되면 미리 짐작할 수 있는 결과가 도출될 가능성이 많으므로 그 가치가 떨어진다. 이러한 문제점을 개선하기 위해 기업에서 중요시하는 전략적「가치속성」과 빈도수를 결합한 전사적인 관점의 연관관계 분석이 필요하다.

전사적 관점에서 전략적 중요도에 대해 고려할 수 있는 요인은 많지만, 본 논문에서는 실증적인 알고리즘의 도출을 위해 4가지 속성을 이용하였다. 이들 속성은 기존에 고려하던 제품의 판매 빈도수(혹은 판매개수) 이외에 해당 제품의 단위당 수익성 및 전략적 마케팅 가치, 고객만족도 등이다. 본 연구에서는 Apriori와 TSAA 알고리즘의 발전된 방법론을 제시하였으며[5] 전략적 중요도를 고려한 항목가중치를 적용하기 위해 새로운 알고리즘을 개발하였다.

2. 전략적 중요도를 고려한 연관규칙 탐사

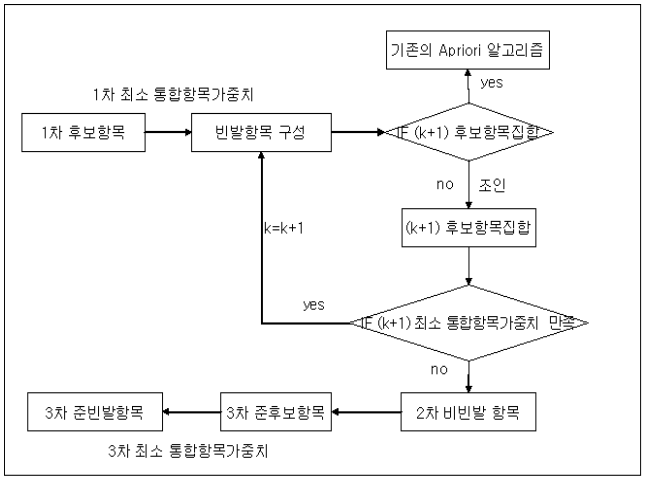
기존의 연구들은 각 항목의 특성을 고려한 가중치나 지지도를 다르게 책정하는 방법들을 주로 사용하고 있다. 항목의 가중치를 사용자의 관심도에 따라 임의로 정하거나[2] 항목의 속성에 가중치를 주어 지지도를 계산하고 새로운 가중지지도를 구하는 것들이 주류였다[1, 3, 4]. 이러한 방법의 문제점은 사용자의 관심 속성에 따라 연관규칙 탐사의 결과가 다르게 나오고, 계산된 항목가중치가 항목의 지지도에 크게 영향을 받아 가중치로서의 의미가 퇴색되어 버린다.

가중치를 적용한 연관규칙의 문헌검토를 통해 제안하는 알고리즘의 초점은 다음과 같다.

- 사용자가 속성에 부여하는 가중치의 주관성을 배제하기 위해 측정 가능한 속성을 선정한다.
- 속성의 가중치가 지지도에 퇴색되지 않도록 독

립적으로 속성의 가중치를 고려하는 방안을 연구한다.

- 유통/판매 기업에서 상품을 판매할 때 전략적 중요성을 지닌 속성을 고려한다.
- 준 빈발항목을 탐사하기 위해 TSAA(Transitive Support Association Apriori)[5]를 참조한다.



(그림 1) 알고리즘의 구조

본 논문은 각 상품의 트랜잭션 데이터, 마케팅 비용과 고객만족도에 대해 가중치를 부여함으로써 연관규칙을 탐사하는 알고리즘을 제안한다. 알고리즘의 전체적인 구성은 (그림 1)과 같다.

1) 기호정의

논문의 알고리즘에 사용되는 기호들은 다음과 같다.

i : 1차 항목, i,j : 2차 항목

π_i : 항목 i에 대한 수익의 정규화된 값

μ_i : 항목 i에 대한 전략적 마케팅가치의 정규화된 값

σ_i : 항목 i에 대한 고객만족도의 정규화된 값

S_i : 항목 i에 대한 빈도수의 정규화된 값(지지도)

p_i : 수익

v_i : 전략적인 마케팅 가치

cs_i : 고객만족도

n : 트랜잭션 데이터베이스에서 발생한 전체 항목수의 합

f_i : 항목 i의 빈도수

w_s : 지지도의 가중치

w_π : 이익의 가중치

w_σ : 고객만족도의 가중치

w_μ : 전략적 마케팅가치의 가중치

R_t : t차 항목의 최소 통합 항목가중치

$p_{i,j}$: 2차 항목집합 {i,j}의 수익

$cs_{i,j}$: 2차 항목집합 {i,j}의 고객만족도

$v_{i,j}$: 2차 항목집합 {i,j}의 전략적 마케팅 가치

$\pi_{i,j}$: 2차 항목집합 {i,j}에 대한 수익의 정규화된 값

$\mu_{i,j}$: 2차 항목집합 {i,j}에 대한 전략적 마케팅 가치의 정규화된 값

$\sigma_{i,j}$: 2차 항목집합 {i,j}에 대한 고객만족도의 정규화된 값

[정의 1] 통합항목가중치

지지도와 각 속성의 정규화된 값과 전략적 중요도를 고려하여 계산한 각 항목의 통합가중치이다.

$$w_i = \sum b_i f_b$$

w_i : 항목 i의 통합가중치,

b_i: 속성 b에 따른 항목 i의 정규화된 값

f_b : 속성 b의 가중치

[정의 2] 최소 통합항목가중치

후보항목 중에서 빈발항목이 되기 위해 만족해야 할 최소 임계값을 말한다.

$$R_t = \frac{1}{t\text{차 후보항목의 수}}$$

2) 알고리즘 수행절차

단계1. 기업에서 상품의 판매를 위해 중요시되는 측정 가능한 속성들을 선정한다.

$$\pi_i = \frac{p_i}{\sum p_i},\ 0 \le \pi_i \le 1,$$

$$\mu_i = \frac{v_i}{\sum v_i},\ 0 \le \mu_i \le 1,$$

$$\sigma_i = \frac{cs_i}{\sum cs_i}\ ,\ 0 \le \sigma_i \le 1 \quad \text{----------식 (1)}$$

단계2. 트랜잭션 데이터베이스를 스캔하여 각 항목의 지지도를 구한다. 기존의 Apriori 알고리즘에서 계산하는 방법과 달리 트랜잭션 데이터베이스에 존재하는 모든 항목의 수를 더한다.

$$S_i = \frac{f_i}{n}\ ,\ 0 \le S_i \le 1 \quad \text{----------식 (2)}$$

단계3. 단계1에서 구한 각각의 정규화된 값과 단계2에서 구한 지지도를 이용하여 각각의 통합항목가중치를 계산한다.

$$w_s + w_\pi + w_\sigma + w_\mu = 1,$$

$$w_i = w_s s_i + w_\pi \pi_i + w_\sigma \sigma_i + w_\mu \mu_i,\ i \in I$$

$$\Omega_i = \frac{\omega_i}{\sum \omega_i} \quad 0 \le w_i \le 1 \quad \text{----------식 (3)}$$

단계4. 지금까지 후보항목의 통합 항목가중치를 계산하였다. 후보항목이 빈발항목으로 되기 위해서는 일정한 값 이상이 되어야 한다. 기존의 Apriori 알고리즘에서는 어떤 항목이 전체 거래 빈도수에서 차지하는 비율로 최소상대지지도를 결정하였다. 본 연구에서는 최소상대지지도를 다른 방법으로 결정한다. 본 연구에서 사용하는 최소상대지지도는 최소 통합 항목가중치(R_t)를 쓰기로 한다.

$$R_1 = \frac{1}{\text{1차 후보항목의 수}} \quad \text{-------식 (4)}$$

1차 후보항목 중에서 1차 최소상대항목가중치(R_1)를 만족하는 1차 빈발항목을 구해야 한다. 단계3에서 구한 항목가중치가 R_1보다 크면 1차 빈발항목이 되고, 그렇지 않은 항목은 제거된다.

단계5. 2차 후보항목의 빈발횟수를 구한 후 모두 합하여 2차 후보항목 각각의 항목가중치를 계산한다.

$$f_{i,j} = f_i + f_j\ ,\ S_{i,j} = \frac{f_{i,j}}{\sum f_{i,j}},\ 0 \le S_{i,j} \le 1$$

---식 (5)

단계6. 2차 후보항목의 통합 항목가중치를 계산하기 전에 단계1과 같이 세 가지 속성에 대해 2차 후보항목의 값을 정규화시켜야 한다.

$$p_{i,j} = p_i + p_j\ ,\ cs_{i,j} = cs_i + cs_j\ ,$$
$$v_{i,j} = v_i + v_j$$

$$\pi_{i,j} = \frac{p_{i,j}}{\sum p_{i,j}},\ \upsilon_{i,j} = \frac{v_{i,j}}{\sum v_{i,j}},$$

$$0 \le \pi_{i,j} \le 1,\ \ 0 \le \mu_{i,j} \le 1,$$

$$\sigma_{i,j} = \frac{cs_{i,j}}{\sum cs_{i,j}},\ \ 0 \le \sigma_{i,j} \le 1 \quad \text{----------식 (6)}$$

단계7. 2차 후보항목의 통합 항목가중치는 단계3과 같은 방식으로 계산한다.

$$w_s + w_\pi + w_\sigma + w_\mu = 1\ ,$$

$$w_{i,j} = w_s s_{i,j} + w_\pi \pi_{i,j} + w_\sigma \sigma_{i,j} + w_\mu \mu_{i,j}$$

$$\Omega_{i,j} = \frac{\omega_{i,j}}{\sum \omega_{i,j}} \quad \text{----------식 (7)}$$

단계8. 2차 후보항목으로부터 2차 빈발항목을 찾기 위해 2차 최소상대항목 가중치를 결정한다.

$$R_2 = \frac{1}{\text{2차 후보항목의 수}} \quad \text{---------- 식 (8)}$$

단계9. 이 단계에서는 2차 빈발항목과 2차 비빈발항목으로 나누어 3차 항목에 대한 탐사를 한다. 2차

빈발항목에 대한 탐사의 결과는 3차 빈발항목집합이 되고, 2차 비빈발항목에 대한 탐사의 결과는 3차 준빈발항목집합이 된다.

3. 적용사례

제시한 알고리즘의 유용성을 검증하기 위하여 실제거래 자료를 사용하여 사례분석을 실시하였다. 자료수집의 대상은 수원에 위치한 H 할인점을 선정하였으며, 고객과의 거래에서 발생한 730여개의 실제 거래자료를 이용하였다

Apriroi, TSAA 및 본 논문의 알고리즘 등 세 가지 알고리즘을 적용하여 분석한 결과를 토대로 의미를 분석한 것이 <표 1>에 나와 있다. 논문의 알고리즘에서는 각 속성별 가중치를 어떻게 설정하였는지, Apriroi 알고리즘과 TSAA 알고리즘에서는 최소지지도와 최소상대지지도를 어떻게 설정하였는지에 따라 빈발항목 집합의 구성이 달라질 수 있다. 위의 사례분석결과를 보면 논문에서 제시한 알고리즘이 다른 알고리즘에 비해 많은 빈발항목을 탐사하는 경향이 있음을 알 수 있다.

<표 1> 알고리즘별 탐사결과 비교

내 용		본 알고리즘	TSAA	Apriroi
개수	1차 빈발항목	9 개	11 개	11 개
	2차 빈발항목	13 개	4 개	4 개
	3차 빈발항목	10 개	5 개	5 개
	4차 빈발항목	9 개	5 개	5 개
	3차 준빈발항목	11 개	11 개	없음
공통점	1차 빈발항목	{4} {6} {8} {15} {16} {20}		
	2차 빈발항목	{4,6} {6,8} {6,15} {6,20}		
	3차 빈발항목	{4,6,8} {4,6,15} {4,6,20} {6,8,15} {6,15,20}		
	4차 빈발항목	{4,6,8,15} {4,6,8,20} {4,6,15,20}{6,8,15,20}		
	3차 준빈발항목	없음		
차이점	1차 빈발항목	{7} {13} {14} {15}	{2} {10} {11} 12} {18}	
	2차 빈발항목	{4,8} {4,13} {4,15} {4,20}{6,13} {8,13} {13,15}{13,20} {15,20}	없음	없음
	3차 빈발항목	{4,6,13} {4,15,20} {6,8,20}{6,13,15} {6,13,20}	없음	없음
	4차 빈발항목	{4,6,13,15}{4,6,13,20} {4,13,15,20}{6,8,13,15} {6,13,15,20}	{4,8,15,20}	
	3차 준빈발항목	전부 포함	전부 포함	

4. 결론

현재까지 알려진 연관규칙 탐사기법은 대부분 빈발항목들만 탐사하게 되어 있는 반면 기업에서 중요하게 생각하는 전략적 속성(수익성, 마케팅가치, 고객만족도 등)을 고려한 연관규칙 탐사기법에 대한 연구는 거의 없었다. 본 논문은 빈발 횟수 이외에 다른 세 가지 속성을 고려하였으며, 각 속성에 전략적 가중치를 반영하였다. 속성별 가중치는 다시 항목의 중요도를 나타내는 항목가중치 계산에 사용되었다. 기존의 가중치를 적용한 연관규칙 알고리즘에서 나타나는 문제점으로 가중치가 지지도에 크게 영향을 받는 등의 문제점이 있었다. 본 연구에서는 정규화를 통하여 그러한 영향을 완화시켜주는 가중치 계산 프로세스를 제시하였다.

세 가지 알고리즘(본 논문, Apriori, TSAA)을 실제 사례 데이터에 적용한 결과 논문의 알고리즘을 통해 탐사된 연관항목 집합은 Apriori 알고리즘으로 얻어지는 연관항목 집합을 거의 모두 포함하였다. 본 알고리즘의 3차 준빈발항목 탐사결과가 TSAA의 결과와 상이하게 나타난 것은 분석의 관점에 따라 다른 각도에서 데이터를 분석할 수 있기 때문이다. 제시한 알고리즘은 유통/판매 기업에서 다른 탐사기법보다 전략적으로 더 의미 있는 연관항목 집합을 탐색하는 데 유용하게 활용할 수 있을 것이다.

참고문헌

[1] Cai, C. H., Ada, W. C., Cheng, C. H. and Kwong, W. W., "Mining Association Rules with Weighted Items," Proc. of 1998 Intl. Database Engineering and Applications Symposium (IDEAS '98), Cardiff, Wales, U.K., pp. 68-77, 1998.

[2] Yue, S., Tsang, E., Yeung, D., Shi, D., "Mining Fuzzy Association Rules with Weighted Items," IEEE International Conference on Systems, Man and Cybernetics, v.3, pp.1906-1911, 2000.

[3] 박종수, 유원경, 홍기형, “연관규칙 탐사와 그 응용,” 정보과학회지 16권, 1998.

[4] 손승현, 김재련, “시간 가중치를 고려한 연관규칙,” 한국산업경영시스템 학회 추계학술대회, 2000.

[5] 현영준, “이행적 연관규칙 탐사기법 연구,” 성균관대학교 석사학위논문(지도교수 최덕원), 2005.