Journal of the Korean Data & Information Science Society
2010, **21**(6), 1147–1154

한국데이터정보과학회지

항목 알에프엠 점수를 고려한 가중 연관성 규칙

박희창[1]

[1]창원대학교 통계학과

접수 2010년 9월 15일, 수정 2010년 10월 26일, 게재확정 2010년 10월 31일

요 약

데이터 마이닝의 중요 목표 중의 하나는 여러 변수들 간의 관계를 발견하고 결정하는 것이다. 이를 위해 필요한 기법인 연관성 규칙은 각 항목들 간의 관련성을 찾아내는 데 활용되며, 지지도, 신뢰도, 향상도 등의 연관성 측도를 기반으로 두 항목간의 관계를 수치화함으로써 의미 있는 규칙을 찾아낸다. 본 논문에서는 수익성이 가장 높은 고객을 찾기 위해 고객 정보를 이용하는 기법으로 가장 널리 사용되어온 방법인 알에프엠 기법을 항목에 적용하여 항목의 알에프엠 점수를 항목의 중요도로 고려하여 가중 연관성 규칙의 평가기준을 제시하였다. 모의실험에서는 일반적인 연관성 규칙과 알에프엠 점수를 가중치로 한 가중 연관성 규칙의 유용성을 비교하였다.

주요용어: 가중 신뢰도, 가중 연관성 규칙, 가중 지지도, 가중 향상도, 데이터마이닝.

1. 서론

데이터마이닝은 대용량의 관측 가능한 데이터를 기반으로 새로운 법칙과 관계, 잠재된 지식, 기대하지 못했던 패턴 등을 발견하고 이를 바탕으로 의사결정을 위한 정보로 활용하고자 하는 것이다. 이 기법은 조직의 최적 전략이나 의사결정을 뒷받침해 줄 수 있는 고급정보가 필요하게 되면서 등장하게 되었다. 데이터마이닝의 대표적인 기법으로는 연관성규칙 (association rule), 의사결정나무 (decision tree) 기법, k-평균 군집방법, 신경망모형 (neural network) 등이 있다.

데이터마이닝 기법 중에서 가장 많이 활용되고 있는 연관성 규칙은 대용량 데이터베이스에서 각 항목들 간의 관련성을 찾아내는 기법으로 여러 가지 연관기준값을 바탕으로 관련성 여부를 측정한다. 이러한 연관성 규칙은 Agrawal 등 (1993)이 최초로 제안하였으며, 그 이후 많은 학자들에 의해 연관성 규칙의 생성에 관한 연구가 수행되었다 (Agrawal과 Srikant, 1994; Park 등 1995; Srikant와 Agrawal, 1995; Toivonen, 1996; Bayardo, 1998; Cai 등, 1998; Han과 Fu, 1999; Liu 등, 1999; Pasquier 등,1999; Han 등, 2000; Pei 등, 2000; Cho와 Park, 2007; Cho와 Park, 2008; Choi와 Park, 2008; Park, 2008).

가중 연관성 규칙 (weighted association rule)은 의사결정자의 중요도의 관점에 따라 개별 항목 각각에 대해 가중치를 부여하여 연관규칙을 찾는 것이다 (이정숙과 김재련, 2003). 가중 연관성 규칙은 기존의 연관성 규칙과는 다르게 아이템 가중값이나 트랜잭션 가중값 등 다양한 가중인자들을 활용하여 후보 아이템 집합을 생성하며, 여러 가중 인자들에 의하여 정의된 최소 가중지지도를 이용하여 빈발 항목 집합을 결정하게 된다. 가중 연관성 규칙을 적용하게 되면 각 항목별로 의사결정자의 입장에서 중요한 속성에 대해서 가중치를 부여하고, 각 항목의 빈발 정도를 고려하여 항목 개별 최소 지지도를 적용하는 것은 단순히 많이 판매되는 항목뿐만이 아니라 항목의 중요도를 고려하여 중요도가 높은 항목에 대해서도

[1] (641-773) 경남 창원시 사림동 9번지, 창원대학교 통계학과, 교수. E-mail: hcpark@changwon.ac.kr

발견할 수 있다. 또한 항목별 빈발 정도를 고려하여 각 항목에 대해서 다른 최소지지도를 적용하면 항목 단위 개수 당 중요도에 대해서도 발견할 수 있다.

RFM (Recency, Frequency, Monetary) 분석기법은 고객의 구매 행동을 구매 시기, 구매 빈도, 구매 금액의 세 가지 척도에 의거하여 각각의 고객들의 기록에 따라 수익성을 평가하는 방법이다 (오윤경 등, 2003). 이러한 RFM 기법은 고객의 미래 잠재가능성을 예측하게 해주며, 앞으로의 판매촉진 비용을 데이터베이스 상의 각각의 고객들로부터 얻을 수 있는 잠재이익과 비교할 수 있도록 해준다.

본 논문에서는 수익성이 가장 높은 고객을 찾기 위해 고객 정보를 이용하는 기법으로 가장 널리 사용되어온 방법인 RFM 기법을 항목에 적용하여 항목의 RFM 점수를 항목의 중요도로 고려하여 가중 연관성 규칙의 평가기준을 제시하고자 한다.

2. 항목 RFM 점수를 이용한 가중 연관성 규칙 평가 기준

일반적인 연관성규칙 탐사의 경우 데이터베이스의 각 트랜잭션을 구성하는 데이터 아이템들은 같은 특성의 항목으로 구성되어있다고 간주한다. 반면에 가중 연관성 규칙 탐사에서는 규칙을 구성하는 임의의 아이템들이 다른 아이템에 비해 더 중요하다면 이에 가중치를 부여하여 규칙으로서 탐사한다는 차이점을 지니고 있다 (Kim 등, 2002). 이러한 가중 연관성 규칙을 생성하기 위한 평가기준을 정의하기 위해 다음과 같은 분할표를 고려한다.

표 2.1 가중 연관성 규칙 생성을 위한 2×2 분할표

		Y		합
		1	0	
X	1	$(w_x + w_y)n_{11}$	$w_x n_{10}$	$w_x(n_{11} + n_{10}) + w_y n_{11}$
	0	$w_y n_{01}$	n_{00}	$w_y n_{01} + n_{00}$
합		$w_x n_{11} + w_y(n_{11} + n_{01})$	$w_x n_{10} + n_{00}$	$w_x(n_{11} + n_{10}) + w_y(n_{11} + n_{01}) + n_{00}$

이 표로부터 가중 연관성 규칙 평가기준을 정의하면 다음과 같다. 먼저 가중 지지도 $supp_w(X \Rightarrow Y)$는 표 2.1로부터 항목 집합 X와 항목 집합 Y가 동시에 발생하는 거래량 (transaction)의 비율을 의미하며, 이를 일반적인 지지도와 구별하기 위해 가중 지지도 (weighted support)라 하고 다음과 같이 정의할 수 있다.

$$supp_w(X \Rightarrow Y) = \frac{(w_x + w_y)n_{11}}{w_x(n_{11} + n_{10}) + w_y(n_{11} + n_{01}) + N_{00}} \tag{2.1}$$

가중 신뢰도 (weighted confidence) $conf_w(X \Rightarrow Y)$는 항목 집합 X가 포함된 거래 비율 중 항목 집합 X와 항목 집합 Y가 동시에 포함된 거래의 비율을 의미하며, 다음과 같이 정의된다.

$$conf_w(X \Rightarrow Y) = \frac{(w_x + w_y)n_{11}}{w_x(n_{11} + n_{10}) + w_y n_{11}} \tag{2.2}$$

가중 향상도 (weighted lift) $lift_w(X \Rightarrow Y)$는 항목 집합 X를 구매한 경우 그 거래가 항목 집합 Y를 포함하는 경우와 항목 집합 Y가 임의로 구매되는 경우의 비를 의미하며, 다음과 같이 정의된다.

$$lift_w(X \Rightarrow Y) = \frac{[(w_x + w_y)n_{11}][w_x(n_{11} + n_{10}) + w_y(n_{11} + n_{01}) + n_{00})]}{[w_x(n_{11} + n_{10}) + w_y n_{11}][w_x n_{11} + w_y(n_{11} + n_{01})]} \tag{2.3}$$

위에서 정의한 식에서 w_x와 w_y는 각각 항목 X와 Y의 가중치를 의미하며, 본 논문에서는 각 항목의 RFM 점수를 가중치로 이용하고자 한다. RFM은 최근성 (recency; R), 최빈성 (frequency; F), 총구매

액 (monetary; M)의 약자로써 세 가지 요소로서 각 개체들의 자료를 점수화하여 고객들을 세분화하는 기법이다. 여기서 최근성은 '고객이 얼마나 최근에 구입했는가?', 최빈성은 '고객이 얼마나 자주 상품을 구입했는가?', 그리고 총구매액은 '고객이 구입한 총구매액은 얼마인가?'를 의미한다. RFM 기법은 이들에 대한 정보를 선형결합에 의해서 축약하고 구입 가능성이 높은 고객들을 세분화시키는 방법으로서 사용하기 간단하고 편리한 모델링 기법으로 알려져 있다 (Lee 등, 2004).일반적으로 RFM의 개념은 R, F, M 변수들 각각에 대해서 고객들에게 점수를 부여하는 것이다. 본 논문에서는 이들을 항목에 대한 최근 판매시기 (R), 총 판매량 (F), 총 판매금액 (M) 등으로 변환하여 식 (2.4)와 같이 선형 결합으로 나타내어 각 항목에 대한 RFM 점수를 부여하고자 한다.

$$RFM\text{점수} = a \times \text{Recency} + b \times \text{Frequency} + c \times \text{Monetary}. \tag{2.4}$$

여기서 a, b, c는 각 변수에 대한 가중치이고, RFM 점수는 각 변수들과 해당 가중치의 선형결합에 의해서 얻어진 점수이다. 고객에 대한 RFM 기법과 마찬가지로 대개의 경우에는 변수 R, F, M 각각에 대해서 분석 자료를 정렬한 후 균등하게 다섯 등분하여 1, 2, 3, 4, 5의 가중치를 부여하며, 가중치를 높게 부여한 집단일수록 중요한 역할을 하는 집단으로 생각한다. 그러나 각 변수들에 대해서 반드시 다섯 등분할 필요는 없고 의사결정자의 목적에 따라 다를 수 있으며, 각 집단에 부여되는 점수 역시 반드시 1, 2, 3, 4, 5로 부여해야 한다는 규칙도 없다. 그러나 많은 연구에서 5개의 집단으로 나누고, 각 집단에 부여되는 점수를 1, 2, 3, 4, 5로 부여하고 있다. 본 연구에서도 R, F, M의 각 변수에 대해서 5개의 집단으로 나누고 5, 4, 3, 2, 1의 점수를 부여하는 방법을 사용한다.

3. 예제를 통한 고찰

본 절에서는 일반적인 연관성 규칙과 RFM 점수를 가중치로 한 가중 연관성 규칙의 유용성을 예제에 의해 비교하고자 한다. 항목 집합 X, Y에 대해 다음과 같이 가정하였다. 먼저 데이터베이스에 있는 총 트랜잭션의 수 (t)를 50명으로 하고, 항목 집합 Y를 결제 방식을 기준으로 신용 카드로 결제 (1)한 사람 수를 30명으로 하고 신용 카드 이외의 방법으로 결제 (0)한 사람의 수를 20명으로 하였다. 항목 집합 X와 Y가 동시에 발생한 빈도 수, 즉 100만원 이상의 냉장고를 구매하면서 신용카드로 결제한 빈도수는 (5+e)명으로 하였다. 또한 항목 집합 X는 구매한 냉장고의 금액을 기준으로 100만원 이상 (1) 구매한 사람 수를 (15+e+r)명으로 하고 100만원 미만 (0)을 구매한 사람 수를 (35-e-r)명으로 하였다. 이를 정리하면 표 3.1과 같다.

표 3.1 모의실험 데이터

		Y		합
		1	0	
X	1	$5+e$	$10+r$	$15+e+r$
	0	$25-e$	$10-r$	$35-e-r$
합		30	20	50

이 표에서 e및 r이 취할 수 있는 정수 값의 범위는 다음과 같다.

$$0 \le e \le 25, 0 \le r \le 10 \tag{3.1}$$

이로부터 e, r, w_x, 그리고 w_y의 변화에 따라 일반적인 연관성 규칙과 가중 연관성 규칙에서의 기본적인 평가기준인 지지도, 신뢰도, 향상도를 계산하여 그 일부를 아래의 표 3.2, 표 3.3, 그리고 표 3.4에

제시하였다. 여기서 $n_{11} = n(X = 1, Y = 1)$, $n_{10} = n(X = 1, Y = 0)$, $n_{01} = n(X = 0, Y = 1)$, $n_{00} = n(X = 0, Y = 0)$을 의미한다.

표 3.2는 항목 X의 가중치 w_x를 1로 하고 항목 Y의 가중치 w_x를 3으로 했을 때의 연관성 규칙 평가기준값을 나타낸 표이다. 모의실험결과를 전체적으로 살펴보면, 일반적인 연관성 규칙을 적용하는 경우에 비해 가중 연관성 규칙을 적용하게 되면 지지도와 신뢰도는 커지는 반면에, 향상도는 대체적으로는 증가하나 간혹 감소하는 경향을 보이기도 한다. 이러한 사실을 좀 더 구체적으로 살펴보기 위해 $n_{11} = 19$, $n_{10} = 13$, $n_{01} = 11$, $n_{00} = 7$인 경우에 지지도와 가중 지지도는 각각 0.380과 0.589로 나타났고, 신뢰도와 가중 신뢰도는 각각 0.594와 0.854로 나타났으며, 향상도 및 가중 향상도는 각각 0.990과 1.011로 계산되었다. 이 결과로부터 알 수 있는 사실은 가중치를 고려한 경우가 훨씬 더 연관성 강도가 강하다는 것이다. 뿐만 아니라 최소지지도를 0.4로 하여 일반적인 연관성 규칙을 탐색하는 경우에는 지지도가 0.4 보다 작아서 규칙이 생성이 되지 않는다. 또한 향상도도 1보다 작게 되어 의미가 없게 된다. 반면에 가중 연관성 규칙을 적용하게 되면 최소지지도의 기준도 만족하고 향상도도 1보다 크게 되는 동시에 신뢰도가 상당히 큰 값을 갖게 되어 연관성 규칙으로 생성된다.

표 3.2 항목 Y의 가중치에 의한 연관성 규칙 평가기준값 ($w_x = 1$, $w_y = 3$)

n_{11}	n_{10}	n_{01}	n_{00}	$supp$	$supp_w$	$conf$	$conf_w$	$lift$	$lift_w$
18	20	12	0	0.360	0.563	0.474	0.783	0.789	0.928
18	19	12	1	0.360	0.563	0.486	0.791	0.811	0.938
18	18	12	2	0.360	0.563	0.500	0.800	0.833	0.948
18	17	12	3	0.360	0.563	0.514	0.809	0.857	0.959
18	16	12	4	0.360	0.563	0.529	0.818	0.882	0.970
18	15	12	5	0.360	0.563	0.545	0.828	0.909	0.981
18	14	12	6	0.360	0.563	0.563	0.837	0.938	0.992
18	13	12	7	0.360	0.563	0.581	0.847	0.968	1.004
18	12	12	8	0.360	0.563	0.600	0.857	1.000	1.016
18	11	12	9	0.360	0.563	0.621	0.867	1.034	1.028
18	10	12	10	0.360	0.563	0.643	0.878	1.071	1.041
19	20	11	0	0.380	0.589	0.487	0.792	0.812	0.937
19	19	11	1	0.380	0.589	0.500	0.800	0.833	0.947
19	18	11	2	0.380	0.589	0.514	0.809	0.856	0.957
19	17	11	3	0.380	0.589	0.528	0.817	0.880	0.967
19	16	11	4	0.380	0.589	0.543	0.826	0.905	0.978
19	15	11	5	0.380	0.589	0.559	0.835	0.931	0.988
19	14	11	6	0.380	0.589	0.576	0.844	0.960	0.999
19	13	11	7	0.380	0.589	0.594	0.854	0.990	1.011
19	12	11	8	0.380	0.589	0.613	0.864	1.022	1.022
19	11	11	9	0.380	0.589	0.633	0.874	1.056	1.034
19	10	11	10	0.380	0.589	0.655	0.884	1.092	1.046

표 3.3은 항목 X의 가중치는 5로 하고 항목 Y의 가중치를 1로 했을 때의 연관성 규칙 평가기준값을 나타낸 표이다. 모의실험결과를 전체적으로 살펴보면, 이 경우에도 일반적인 연관성 규칙을 적용하는 경우에 비해 가중 연관성 규칙을 적용하게 되면 지지도와 신뢰도는 커지는 반면에, 향상도는 대체적으로는 증가하나 간혹 감소하는 경향을 보이기도 한다. 이를 확인하기 위해 위와 마찬가지로 $n_{11} = 19$, $n_{10} = 13$, $n_{01} = 11$, $n_{00} = 7$인 경우를 자세히 살펴보면 지지도와 가중 지지도는 각각 0.380과 0.579로 나타났고, 신뢰도와 가중 신뢰도는 각각 0.594와 0.637로 나타났으며, 향상도 및 가중 향상도는 각각 0.990과 1.004로 계산되었다. 이 결과로부터도 알 수 있는 사실은 가중치를 고려한 경우가 더 연관성 강도가 강하다는 것이다. 또한 최소지지도가 0.4이면 일반적인 연관성 규칙을 탐색하는 경우에는 지지도가 0.4 보다 작아서 규칙이 생성이 되지 않는다. 또한 향상도도 1보다 작게 되어 의미가 없게 된다.

이 경우에도 가중 연관성 규칙을 적용하게 되면 최소지지도의 기준도 만족하고 향상도도 1보다 크게 되는 동시에 신뢰도가 큰 값을 갖게 되어 연관성 규칙으로 생성된다. 표 3.2와 표 3.3을 비교해보면 항목 Y의 가중치가 항목 X의 가중치에 비해 상대적으로 크게 되면 연관성 평가기준값이 커지는 것을 알 수 있다.

표 3.3 항목 X의 가중치에 따른 연관성 규칙 평가기준값 ($w_x = 5$, $w_y = 1$)

n_{11}	n_{10}	n_{01}	n_{00}	$supp$	$supp_w$	$conf$	$conf_w$	$lift$	$lift_w$
18	20	12	0	0.360	0.491	0.474	0.519	0.789	0.952
18	19	12	1	0.360	0.500	0.486	0.532	0.811	0.958
18	18	12	2	0.360	0.509	0.500	0.545	0.833	0.964
18	17	12	3	0.360	0.519	0.514	0.560	0.857	0.970
18	16	12	4	0.360	0.529	0.529	0.574	0.882	0.977
18	15	12	5	0.360	0.540	0.545	0.590	0.909	0.984
18	14	12	6	0.360	0.551	0.563	0.607	0.938	0.991
18	13	12	7	0.360	0.563	0.581	0.624	0.968	0.999
18	12	12	8	0.360	0.574	0.600	0.643	1.000	1.007
18	11	12	9	0.360	0.587	0.621	0.663	1.034	1.016
18	10	12	10	0.360	0.600	0.643	0.684	1.071	1.025
19	20	11	0	0.380	0.507	0.487	0.533	0.812	0.959
19	19	11	1	0.380	0.516	0.500	0.545	0.833	0.964
19	18	11	2	0.380	0.525	0.514	0.559	0.856	0.970
19	17	11	3	0.380	0.535	0.528	0.573	0.880	0.976
19	16	11	4	0.380	0.545	0.543	0.588	0.905	0.983
19	15	11	5	0.380	0.556	0.559	0.603	0.931	0.989
19	14	11	6	0.380	0.567	0.576	0.620	0.960	0.996
19	13	11	7	0.380	0.579	0.594	0.637	0.990	1.004
19	12	11	8	0.380	0.591	0.613	0.655	1.022	1.012
19	11	11	9	0.380	0.603	0.633	0.675	1.056	1.020
19	10	11	10	0.380	0.616	0.655	0.695	1.092	1.029

표 3.4는 w_x를 5로 하고 w_y를 3으로 했을 때의 연관성 규칙 평가기준값을 나타낸 표이다. 모의실험 결과를 전체적으로 살펴보면, 위의 경우와 마찬가지로 일반적인 연관성 규칙을 적용하는 경우에 비해 가중 연관성 규칙을 적용하게 되면 지지도와 신뢰도는 커지는 반면에, 향상도는 대체적으로는 증가하나 간혹 감소하는 경향을 보이기도 한다. $n_{11} = 19$, $n_{10} = 13$, $n_{01} = 11$, $n_{00} = 7$인 경우에 지지도와 가중 지지도는 각각 0.380과 0.591로 나타났고, 신뢰도와 가중 신뢰도는 각각 0.594와 0.700으로 나타났으며, 향상도 및 가중 향상도는 각각 0.990과 0.973으로 계산되었다. 이 결과로부터 알 수 있는 사실은 가중치를 고려한 경우가 훨씬 더 연관성 강도가 강하다는 것이다. 뿐만 아니라 최소지지도가 0.4이면 일반적인 연관성 규칙을 탐색하는 경우에는 지지도가 0.4 보다 작아서 규칙이 생성이 되지 않는다. 반면에 가중 연관성 규칙을 적용하게 되면 최소지지도의 기준도 만족하고 신뢰도가 큰 값을 갖게 되어 연관성 규칙으로 생성하게 된다. 표에는 나타내지는 못했으나 향상도의 값이 1보다 큰 경우에는 가중 향상도도 1보다 크게 되는 사실을 확인할 수 있었다.

표 3.4 항목 X, Y의 가중치에 따른 연관성 규칙 평가기준값 ($w_x = 5$, $w_y = 3$)

n_{11}	n_{10}	n_{01}	n_{00}	$supp$	$supp_w$	$conf$	$conf_w$	$lift$	$lift_w$
18	20	12	0	0.360	0.514	0.474	0.590	0.789	0.918
18	19	12	1	0.360	0.522	0.486	0.603	0.811	0.924
18	18	12	2	0.360	0.529	0.500	0.615	0.833	0.930
18	17	12	3	0.360	0.537	0.514	0.629	0.857	0.936
18	16	12	4	0.360	0.545	0.529	0.643	0.882	0.943
18	15	12	5	0.360	0.554	0.545	0.658	0.909	0.950
18	14	12	6	0.360	0.563	0.563	0.673	0.938	0.957
18	13	12	7	0.360	0.571	0.581	0.689	0.968	0.965
18	12	12	8	0.360	0.581	0.600	0.706	1.000	0.973
18	11	12	9	0.360	0.590	0.621	0.724	1.034	0.981
18	10	12	10	0.360	0.600	0.643	0.742	1.071	0.990
19	20	11	0	0.380	0.533	0.487	0.603	0.812	0.929
19	19	11	1	0.380	0.541	0.500	0.615	0.833	0.935
19	18	11	2	0.380	0.549	0.514	0.628	0.856	0.940
19	17	11	3	0.380	0.557	0.528	0.641	0.880	0.946
19	16	11	4	0.380	0.565	0.543	0.655	0.905	0.953
19	15	11	5	0.380	0.574	0.559	0.670	0.931	0.959
19	14	11	6	0.380	0.582	0.576	0.685	0.960	0.966
19	13	11	7	0.380	0.591	0.594	0.700	0.990	0.973
19	12	11	8	0.380	0.601	0.613	0.717	1.022	0.981
19	11	11	9	0.380	0.610	0.633	0.734	1.056	0.988
19	10	11	10	0.380	0.620	0.655	0.752	1.092	0.997

4. 결론

가중 연관성 규칙은 항목별로 분석가의 관점에서 가중치를 부여하고 각 항목의 빈발 정도를 고려하여 연관규칙을 찾는 것이다. 이러한 가중 연관성 규칙은 다양한 가중인자들을 활용하여 후보 아이템 집합을 생성하며, 정의된 최소 가중지지도를 이용하여 빈발 항목 집합을 결정하게 된다.

본 논문에서는 수익성이 가장 높은 고객을 찾기 위해 고객 정보를 이용하는 기법으로 가장 널리 사용되어온 방법인 RFM 기법을 항목에 적용하여 항목의 RFM 점수를 항목의 중요도로 고려하여 가중 연관성 규칙의 평가기준을 제시하였다. 이 RFM 분석기법은 최근 구매 시기, 구매 빈도, 구매 금액의 세 가지 척도에 의거하여 각각의 고객들의 기록에 따라 수익성을 평가하는 방법으로 고객의 미래 잠재가능성을 예측하게 해준다. 또한 모의실험을 통하여 일반적인 연관성 규칙과 여러 가지 RFM 점수를 가중치로 한 가중 연관성 규칙의 유용성을 비교하였다. 그 결과, 일반적인 연관성 규칙을 적용하는 경우에 비해 지지도와 신뢰도는 커지는 반면에, 향상도는 대체적으로는 증가하나 간혹 감소하는 경향을 보이기도 하였으며, 가중 연관성 규칙을 적용한 경우가 훨씬 더 연관성 강도가 강한 것으로 나타났다. 추가적으로 알 수 있었던 사실은 하나의 항목 가중치가 다른 하나의 항목 가중치에 비해 상대적으로 크게 되면 연관성 평가기준값이 커진다는 것이다. 향후에는 보다 다양한 가중치를 찾아서 이에 대한 가중 연관성 규칙에 관한 연구가 필요할 것으로 사료된다.

참고문헌

오윤경, 김지경, 김상훈 (2003). 고객정보의 종류와 양이 구매모형 예측력에 미치는 영향에 관한 연구. <경영논집>, **37**, 91-121.

이정숙, 김재련 (2003). 항목별 최소지지도와 가중 항목을 고려한 연관규칙. <한국산업경영시스템학회 2003 추계학술대회논문집>, 31-35.

Agrawal, R., Imielinski, R. and Swami, A. (1993). Mining association rules between sets of items in large databases. *Proceedings of the ACM SIGMOD Conference on Management of Data*, 207-216.

Agrawal, R. and Srikant, R. (1994). Fast algorithms for mining association rules. *Proceedings of the 20th VLDB Conference*, 487-499.

Bayardo, R. J. (1998). Efficiently mining long patterns from databases. *Proc. of ACM SIGMOD Conference on Management of Data*, 85-93.

Cai, C. H., Fu, A. W. C., Cheng, C. H. and Kwong, W. W. (1998). Mining association rules with weighted items. *Proceedings of International Database Engineering and Applications Symposium*, 68-77.

Cho, K. H. and Park, H. C. (2007). Association rule mining by environmental data fusion. *Journal of the Korean Data & Information Science Society*, **18**, 279-287.

Cho, K. H. and Park, H. C. (2008). A study of association rule application using self-organizing map for fused data. *Journal of the Korean Data & Information Science Society*, **19**, 95-104.

Choi, J. H. and Park, H. C. (2008). Comparative study of quantitative data binning methods in association rule. *Journal of the Korean Data & Information Science Society*, **19**, 903-910.

Han, J. and Fu, Y. (1999). Mining multiple-level association rules in large databases. *IEEE Transactions on Knowledge and Data Engineering*, **11**, 68-77.

Han, J., Pei, J. and Yin, Y. (2000). Mining frequent patterns without candidate generation. *Proceedings of ACM SIGMOD Conference on Management of Data*, 1-12.

Kim, J., Ceong, H. and Won, Y. (2002). Weighted association rule mining for item groups with different properties and risk assessment for networked systems. *IEICE Transaction on Information and Systems*, **85**, 1-7.

Lee, S., Choi, S., Kim, K. and Kang, C. (2004). Study on development the optimal RFM model for customer segmentation. *Journal of the Korean Data Analysis SocietyI*, **6**, 1829-1840.

Liu, B., Hsu, W. and Ma, Y. (1999). Mining association rules with multiple minimum supports. *Proceedings of the 5th International Conference on Knowledge Discovery and Data Mining*, 337-241.

Park, H. C. (2008). The proposition of conditionally pure confidence in association rule mining. *Journal of the Korean Data & Information Science Society*, **19**, 1141-1151.

Park J. S., Chen M. S. and Philip S. Y. (1995). An effective hash-based algorithms for mining association rules. *Proceedings of ACM SIGMOD Conference on Management of Data*, 175-186.

Pasquier, N., Bastide, Y., Taouil, R. and Lakhal, L. (1999). Discovering frequent closed itemsets for association rules. *Proceedings of the 7th International Conference on Database Theory*, 398-416.

Pei, J., Han, J. and Mao, R. (2000). CLOSET: An efficient algorithm for mining frequent closed itemsets. *Proceedings of ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, 21-30.

Srikant, R. and Agrawal, R. (1995). Mining generalized association rules. *Proceedings of the 21st VLDB Conference*, 407-419.

Toivonen, H. (1996). Sampling large database for association rules. *Proceedings of the 22nd VLDB Conference*, 134-145.

Journal of the Korean Data & Information Science Society
2010, **21**(6), 1147–1154

한국데이터정보과학회지

Weighted association rules considering item RFM scores

Hee Chang Park[1]

[1]Department of Statistics, Changwon National University

Received 15 September 2010, revised 26 October 2010, accepted 31 October 2010

Abstract

One of the important goals in data mining is to discover and decide the relationships between different variables. Association rules are required for this technique and it find meaningful rules by quantifying the relationship between two items based on association measures such as support, confidence, and lift. In this paper, we presented the evaluation criteria of weighted association rule considering item RFM scores as importance of items. Original RFM technique has been used most widely applied method using customer information to find the most profitable customers. And then we compared general association rule technique with weighted association rule technique through the simulation data.

Keywords: Data mining, weighted association rule, weighted confidence, weighted lift, weighted support.

[1] Professor, Department of Statistics, Changwon National University, Changwon, Kyungnam 641-773, Korea. E-mail: hcpark@sarim.changwon.ac.kr