Journal of the Korean Data & Information Science Society
2012, **23**(4), 717–725

http://dx.doi.org/10.7465/jkdi.2012.23.4.717
한국데이터정보과학회지

관심 항목의 발생 가능한 규칙의 수를 고려한 연관성 평가기준

박희창[1]

[1]창원대학교 통계학과

접수 2012년 6월 12일, 수정 2012년 6월 26일, 게재확정 2012년 7월 8일

요약

데이터 마이닝은 데이터베이스로부터 쉽게 드러나지 않는 의미 있는 정보를 생성하는 기법이다. 이 중에서 연관성 규칙은 일반적으로 발생 여부를 나타내는 자료를 이용하여 지지도, 신뢰도, 향상도 등을 수치화함으로써 항목들 간의 관련성을 나타낸다. 기존의 연관성 규칙은 발생 빈도의 크기를 고려하지 않음으로써 정보 손실에 의한 오류를 범할 수 있다. 이를 위해 본 논문에서는 발생 가능한 규칙의 수를 고려한 연관성 평가 기준들을 제안하고 예제를 통하여 기존 연구와 비교한 후, 본 논문에서 제안한 연관성 평가 기준의 유용성을 살펴보았다. 실제 데이터를 통하여 분석한 결과, 기존의 연관성 규칙 평가 기준은 관심항목 수와 트랜잭션의 수를 2배로 하여도 지지도와 신뢰도, 향상도의 값이 동일한 반면에 본 논문에서 제안한 평가 기준은 발생 가능한 규칙의 수를 고려하기 때문에 각각의 평가 기준의 값들이 트랜잭션의 수에 따라 다르다는 것을 알 수 있었다. 또한 본 논문에서 제안하는 평가 기준이 기존의 연관성 규칙 평가 기준에 비해 좀 더 정확한 정보를 제공하는 것을 알 수 있다. 특히 본 논문에서 제안한 신뢰도의 범위가 기존 연관성 평가 기준에 비해 크므로 좀 더 비교 가능한 정보를 제공하는 동시에 향상도의 비교를 용이하게 한다고 할 수 있다.

주요용어: 발생 가능한 규칙의 수, 신뢰도, 연관성 규칙, 지지도, 향상도.

1. 서론

오늘날 현업에서는 정보 기술 및 데이터베이스를 통한 비즈니스 기회 창출 및 확대를 위해 다양한 데이터 마이닝 기법들을 개발하여 활용하고 있다. 데이터 마이닝 기법은 데이터베이스로부터 쉽게 드러나지 않는 의미 있는 정보를 발굴하는 것으로, 이에는 연관성 규칙, 의사결정나무, 군집분석, 신경망 등이 있다. 이들 중에서 연관성 규칙 마이닝 (association rule mining) 방법은 대용량 데이터베이스에 내재되어 있는 항목들 간의 관련성을 찾아내는 데 활용되고 있으며, 항목들 간의 관계를 수치화하여 이들 간의 관련성을 표시함으로써 제조업, 유통업, 그리고 의료분야 등 여러 분야에서 다양하게 적용되고 있다 (Park, 2010b).

Agrawal 등 (1993)에 의해 처음 소개된 연관성 규칙은 이후 국내외적으로 많은 학자들에 의해 연구되어 왔으며, 지금도 활발한 연구가 진행되고 있다 (Agrawal과 Srikant, 1994; Park 등, 1995; Toivonen, 1996; Cai 등, 1998; Han과 Fu, 1999; Liu 등, 1999; Pasquier 등, 1999; Han 등, 2000; Pei 등, 2000; Choi와 Park, 2008; Cho와 Park, 2008; Park, 2010a, 2010b, 2010c; Park, 2011a).

의미 있는 연관성 규칙을 탐색하기 위한 흥미도 측도에는 지지도 (support), 신뢰도 (confidence), 향상도 (lift) 등이 있으며, 일반적인 연관성 규칙 생성과정은 먼저 사용자가 지정한 최소 지지도를

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

만족시키는 빈발항목집합을 생성한 후, 이들에 대해 최저신뢰도 기준을 만족하고 향상도가 1 이상인 것을 규칙으로 채택하게 된다. 이와 같은 기존의 연관성 규칙 마이닝은 발생 유무만을 고려하여 규칙을 생성하여 왔으며, 이는 발생 빈도의 크기를 고려하지 않음으로써 정보 손실에 의한 오류를 범하거나 세밀하지 못한 해석을 할 수도 있다.

이러한 문제를 해결하기 위해 발생빈도가 적은 경우에 대해서는 순위결정함수를 통해 여러 학자들(Wu 등, 2004; Park, 2010a; Park, 2010b)이 해결방안을 제시해왔지만 발생빈도가 많은 경우에는 아직 언급된 바가 없다. 이러한 문제를 해결하기 위해 Lim 등 (2010)은 마케팅 분야에서의 구매건수 관점에서 발생빈도를 고려한 새로운 연관성 규칙을 제시한 후, 이를 한국프로야구 타자들의 성적에 적용한 바 있다. 이 연구는 양적인 의미를 내포하는 경우에 한 케이스 내의 관심변수의 발생빈도를 고려한 가중치를 부여하여 기존의 문제를 해결한 것이다. 하지만 Lim 등 (2010)이 제시한 연관성 평가 기준들은 일반적인 평가 기준값에 비해 상당히 작은 값으로 나타나고 있어서 기존의 방법에 익숙한 사용자들에게는 혼란을 초래할 수 있다. 이를 위해 본 논문에서는 발생 가능한 규칙의 수를 고려한 연관성 평가 기준들을 제안하고 실제 예제를 통하여 기존 연구와 비교한 후, 본 논문에서 제안한 연관성 평가 기준의 유용성을 살펴보고자 한다.

2. 발생 가능한 규칙의 수를 고려한 연관성 규칙

본 절에서는 관심 항목의 발생 가능한 규칙의 수를 고려한 연관성 규칙을 제안하고자 한다. 먼저 기존의 연관성 규칙의 평가기준인 지지도, 신뢰도, 향상도 등을 수식으로 나타내기 위해 다음과 같은 분할표를 고려하기로 한다 (Park, 2011b).

Table 2.1 2×2 contingency table

		Y		Total
		1	0	
X	1	n_{11}	n_{10}	$n_{1.}$
	0	n_{01}	n_{00}	$n_{0.}$
Total		$n_{.1}$	$n_{.0}$	n

지지도 $S(X \Rightarrow Y)$는 항목 집합 X와 항목 집합 Y가 동시에 발생하는 거래의 비율을 의미하며, 신뢰도 $C(X \Rightarrow Y)$는 항목 집합 X가 포함된 거래 비율 중 항목 집합 X와 항목 집합 Y가 동시에 포함된 거래의 비율을 의미하며, 향상도 $L(X \Rightarrow Y)$는 항목 집합 X를 구매한 경우 그 거래가 항목 집합 Y를 포함하는 경우와 항목 집합 Y가 임의로 구매되는 경우의 비를 의미하며, 이들은 다음과 같이 정의된다.

$$S(X \Rightarrow Y) = P(X \cap Y) = \frac{n_{11}}{n}$$

$$C(X \Rightarrow Y) = P(Y|X) = \frac{n_{11}}{n_{1.}}$$

$$L(X \Rightarrow Y) = \frac{P(Y|X)}{P(Y)} = \frac{P(X \cap Y)}{P(X) \cdot P(Y)} = \frac{n_{11} \cdot n}{n_{1.} \cdot n_{.1}}$$

위의 세 가지 수식에서 나타난 바와 같이 지지도, 신뢰도, 향상도는 서로 밀접한 관계를 가지고 있으며, 그 관계성은 쉽게 이해할 수 있다. 이러한 기존의 연관성 규칙 평가 기준이 현업에서 일반적으로 많이 적용되어 왔으나, Lim 등 (2010)이 지적한 바와 같이 이 측도들은 어떤 특정한 사건의 발생 여부만을 기준으로 하여 계산되고 있다. 다시 말하면 이들은 특정한 트랜잭션 내에서 특정 사건이 몇 번 일어났는지에 대해서는 고려하지 않음으로써 정보의 손실을 가져온다고 할 수 있다. Lim 등은

이를 보완하기 위해 발생 빈도를 고려한 연관성 평가 기준을 제안한 바 있으며, 이들의 연구 결과는 기존의 연관성 규칙에서 보다 더 세밀하게 데이터의 상황을 반영한다고 볼 수 있다. 그러나 연관성 평가 기준의 지지도와 신뢰도의 값이 매우 작은 값으로 나타나고 있어서 일반적으로 알려진 최저 지지도와 최저 신뢰도의 값보다 항상 값이 작아서 현업에 적용할 경우 혼란을 초래할 수도 있다. 이러한 문제점을 해결하기 위해 본 논문에서는 관심 항목의 발생 가능한 규칙의 수를 고려한 지지도 $S_P(X \Rightarrow Y)$, 신뢰도 $C_P(X \Rightarrow Y)$, 그리고 향상도 $L_P(X \Rightarrow Y)$를 연관성 평가 기준으로 제안한다.

$$S_P(X \Rightarrow Y) = \frac{\sum_{i=1}^{n} \frac{\#(X=1,\ Y=1)}{\text{발생항목수}}}{\text{전체거래발생수}}$$

$$C_P(X \Rightarrow Y) = \frac{\sum_{i=1}^{n} \frac{\#(X=1,\ Y=1)}{\text{발생항목수}}}{\sum_{i=1}^{n} \frac{\#(X=1,\ Y=1)}{\text{발생항목수}} + \sum_{i=1}^{n} \frac{\#(X=1,\ Y=0)}{\text{발생항목수}}}$$

$$L_P(X \Rightarrow Y) = \frac{\text{전체거래발생수} \times \sum_{i=1}^{n} \frac{\#(X=1,\ Y=1)}{\text{발생항목수}}}{\sum_{i=1}^{n} \frac{\#(X=1)}{\text{발생항목수}} \times \sum_{i=1}^{n} \frac{\#(Y=1)}{\text{발생항목수}}}$$

여기서 발생항목수는 각 케이스 내에서 발생한 항목의 수, $\#(X=1,\ Y=1)$은 항목 X와 Y가 동시에 발생한 경우에 포함된 발생 가능한 규칙의 수, $\#(X=1,\ Y=0)$은 항목 X는 발생하고 Y는 발생하지 않은 경우에 포함된 항목 X의 개수를 의미한다. 그리고 $\#(X=1)$는 $\#(X=1,\ Y=1)$와 $\#(X=1,\ Y=0)$의 합이고, $\#(Y=1)$는 $\#(X=1,\ Y=1)$와 $\#(X=0,\ Y=1)$의 합을 의미한다.

기존의 연관성 규칙 평가 기준과 Lim 등의 연구 결과, 그리고 본 논문에서 제안하는 평가 기준을 비교하기 위해 Table 2.2와 같은 예제 데이터를 이용하고자 한다.

Table 2.2 Sample data(1)

data 1-1	data 1-2	data 1-3	data 1-4	data 1-5
XXXY	XXYY	XYYY	XYZ	XYZ
XXXY	XXYY	XYYY	XYZ	XYZ
XXXY	XXYY	XYYY	XXYZ	XYZ
XXXY	XXYY	XYYY	XZZ	XZZ
XXXY	XXYY	XYYY	XXYYZ	XYZ

이 표의 각 데이터로부터 기존의 연관성 규칙 평가 기준 (general), Lim 등의 연구 결과 (Lim), 그리고 본 논문에서 제안하는 평가 기준 (Park)을 계산한 결과는 다음과 같다.

Table 2.3 Association rule thresholds by sample data(1)

method	threshold	data 1-1	data 1-2	data 1-3	data 1-4	data 1-5
general	support	1.0	1.0	1.0	0.8	0.8
	confidence	1.0	1.0	1.0	0.8	0.8
	lift	1.0	1.0	1.0	1.0	1.0
Lim	support	0.75	1.0	0.75	0.4058	0.356
	confidence	0.5	1.0	1.5	0.534	0.533
	lift	1.0	1.0	1.0	1.014	1.0
Park	support	0.75	1.0	0.75	0.393	0.267
	confidence	1.0	1.0	1.0	0.855	0.8
	lift	1.333	1.0	1.333	2.201	3.0

이 표에서 data 1-1과 data 1-2를 비교해보면 기존의 연관성 규칙 평가 기준에 의하면 모든 평가기준이 1의 값을 취하고 있는 반면에 Lim 등의 연구 결과와 본 논문에서 제안하는 평가 기준은 신뢰도와 지지도, 또는 지지도와 향상도의 값이 두 데이터 간에 다르다는 것을 알 수 있다. 또한 data 1-4와 data 1-5를 비교해보면 기존의 연관성 규칙 평가 기준이 동일한 반면에 Lim 등의 연구 결과와 본 논문에서 제안하는 평가 기준은 평가 기준 모두가 다르다는 것을 알 수 있다. 따라서 Lim 등의 연구 결과와 본 논문에서 제안하는 평가 기준은 기존의 연관성 규칙 평가 기준에 비해 비교 가능한 정보를 제공할 수 있다고 생각된다.

이번에는 각 방법들에 대한 비교를 좀 더 구체적으로 하기 위해 다음의 데이터를 활용하고자 한다. 여기서 data 2-2는 data 2-1의 각 케이스 내에서 발생한 항목의 수를 2배로 한 것이다.

Table 2.4 Sample data(2)

data 2-1	data 2-2
XYABCD	XYABCDXYABCD
XXYABC	XXYABCXXYABC
XABCDE	XABCDEXABCDE
XYYABCDEEE	XYYABCDEEEXYYABCDEEE
YAAAABBBCC	YAAAABBBCCYAAAABBBCC

이 표로부터 기존의 연관성 규칙 평가 기준과 Lim 등의 연구 결과, 그리고 본 논문에서 제안하는 평가 기준을 계산한 결과는 Table 2.5와 같다.

Table 2.5 Association rule thresholds by sample data(2)

method	threshold	data 2-1	data 2-2
general	support	0.600	0.600
	confidence	0.750	0.750
	lift	0.937	0.937
Lim	support	0.083	0.083
	confidence	0.270	0.270
	lift	1.064	1.064
Park	support	0.140	0.700
	confidence	0.801	0.894
	lift	7.208	14.416

기존의 연관성 규칙 평가 기준과 Lim 등의 연구 결과는 관심항목 수와 트랜잭션의 수를 2배로 하여도 지지도와 신뢰도, 향상도의 값은 동일하다. 그러나 본 논문에서 제안한 평가 기준은 발생 가능한 규칙의 수를 고려하기 때문에 각각의 평가 기준의 값들이 트랜잭션의 수에 따라 달라지고 있다. 따라서 본 논문에서 제안한 각 평가 기준이 기존의 연관성 규칙은 물론이고 Lim 등의 연구 결과에 비해 비교 가능한 정보를 제공하는 것을 알 수 있다.

3. 예제를 통한 고찰

본 절에서는 관심 항목의 발생 가능한 규칙의 수를 고려한 연관성 평가 기준의 유용성을 살펴보기 위해 2011년도 한국프로야구 롯데 자이언트 1군 선수들 중에서 규정타석을 채운 선수들을 대상으로 각 방법에 따른 연관성 평가 기준을 미니탭 16에 의해 계산하였다. 먼저 본 논문에서 제안한 지지도의 크기순으로 상위 25개의 규칙을 정렬한 결과를 Table 3.1에 제시하였다. 여기서 지지도는 고려하는

두 선수가 동시에 안타를 치는 비율을 의미한다. 이 표에서 LEE는 이대호, HONG는 홍성흔, KIM은 김주찬, CHO는 조성환, SON은 손아섭, KANG는 강민호, 그리고 GAR은 가르시아를 의미한다.

Table 3.1 Association rule thresholds sorted by support value of Park's method

player	support			confidence			lift		
	general	Lim	Park	general	Lim	Park	general	Lim	Park
LEE ⇒ HONG	0.6095	0.0719	0.1816	0.7619	0.2687	0.8608	0.9876	0.9943	4.0196
HONG ⇒ LEE	0.6095	0.0719	0.1816	0.7901	0.2661	0.8481	0.9876	0.9943	4.0196
KIM ⇒ HONG	0.5894	0.0639	0.1616	0.8115	0.2787	0.8682	1.0280	0.9463	4.1982
HONG ⇒ KIM	0.5894	0.0639	0.1616	0.7466	0.2171	0.7814	1.0280	0.9463	4.1982
LEE ⇒ CHO	0.6074	0.0609	0.1597	0.7471	0.2263	0.7920	0.9992	0.8805	4.2420
CHO ⇒ LEE	0.6074	0.0609	0.1597	0.8125	0.2370	0.8554	0.9992	0.8805	4.2420
CHO ⇒ HONG	0.5652	0.0635	0.1588	0.7647	0.2422	0.8036	0.9771	0.8820	3.9269
HONG ⇒ CHO	0.5652	0.0635	0.1588	0.7222	0.2315	0.7761	0.9771	0.8820	3.9269
SON ⇒ LEE	0.5625	0.0596	0.1489	0.8076	0.2693	0.8697	1.0279	1.0384	4.6417
LEE ⇒ SON	0.5625	0.0596	0.1489	0.7159	0.2301	0.7947	1.0279	1.0384	4.6417
SON ⇒ HONG	0.5510	0.0596	0.1448	0.7941	0.2824	0.8562	0.9977	0.9968	4.5564
HONG ⇒ SON	0.5510	0.0596	0.1448	0.6923	0.2104	0.7706	0.9977	0.9968	4.5564
KIM ⇒ LEE	0.5495	0.0549	0.1444	0.8026	0.2477	0.8490	1.0124	0.9570	4.5568
LEE ⇒ KIM	0.5495	0.0549	0.1444	0.6931	0.2123	0.7753	1.0124	0.9570	4.5568
KANG ⇒ CHO	0.5319	0.0512	0.1317	0.7812	0.2477	0.8383	1.0343	0.9384	4.6925
CHO ⇒ KANG	0.5319	0.0512	0.1317	0.7042	0.1941	0.7375	1.0343	0.9384	4.6925
KANG ⇒ LEE	0.5675	0.0491	0.1316	0.8076	0.2324	0.8500	0.9961	0.8925	4.8045
LEE ⇒ KANG	0.5675	0.0491	0.1316	0.7000	0.1887	0.7442	0.9961	0.8925	4.8045
KIM ⇒ CHO	0.5368	0.0509	0.1299	0.7500	0.2138	0.7881	1.0178	0.8477	4.5372
CHO ⇒ KIM	0.5368	0.0509	0.1299	0.7285	0.2019	0.7478	1.0178	0.8477	4.5372
KANG ⇒ HONG	0.5368	0.0481	0.1294	0.7727	0.2269	0.8006	1.0195	0.8958	4.7174
HONG ⇒ KANG	0.5368	0.0481	0.1294	0.7083	0.1902	0.7625	1.0195	0.8958	4.7174
GAR ⇒ HONG	0.4848	0.0438	0.1159	0.8275	0.2452	0.8509	1.0503	0.8898	4.8345
HONG ⇒ GAR	0.4848	0.0438	0.1159	0.6153	0.1589	0.6584	1.0503	0.8898	4.8345
GAR ⇒ CHO	0.4226	0.0440	0.1138	0.7321	0.2486	0.8202	0.9863	0.9527	4.7837

이 표에서 보는 바와 같이 Lim 등의 연구 결과와 본 논문에서 제시하는 지지도가 기존의 지지도에 비해 좀 더 비교 가능한 정보를 제공하고 있다. 이를 좀 더 구체적으로 설명하면 KIM ⇒ CHO, CHO ⇒ KIM, KANG ⇒ HONG, HONG ⇒ KANG의 규칙에서 기존의 지지도는 모두 0.5368로 동일하나, Lim 등의 연구에서의 지지도는 0.0509와 0.0481이고, 우리의 지지도는 0.1299와 0.1294로 다른 값을 갖는 것으로 계산되었다. 또한 각 방법에 의한 지지도의 변화를 살펴보면 기존의 연관성 규칙에서는 최소 지지도와 최대 지지도가 각각 0.3861과 0.6095로 나타나서 그 범위가 0.2234로 나타났으며, Lim 등의 연구에서의 최소 지지도와 최대 지지도는 각각 0.0341과 0.0719로 나타나서 범위가 0.0378이다. 또한 본 논문에서 제시한 지지도의 최소값과 최대값은 각각 0.0838과 0.1816으로 그 범위는 0.0978로 나타났다. 이로부터 알 수 있는 바와 같이 본 논문에서 제시한 지지도의 범위와 Lim 등의 연구에서의 지지도의 범위가 기존의 연관성 규칙에서의 지지도보다 작게 나타나고 있으나, 본 논문에서 제시한 지지도의 범위가 Lim 등의 연구에서의 지지도의 범위 보다 크므로 Lim 등의 연구 결과보다는 지지도 비교를 용이하게 한다고 할 수 있다. 한편, Lim 등의 연구와 본 논문에서 제안한 지지도는 둘 다 기존값과 많이 차이가 나는 것을 알 수 있는데, 그 이유는 두 공식 모두 발생항목수로 나눈 후 이 값을 다시 전체거래발생수로 나눔으로써 기존의 지지도에 비해 작은 값으로 나타났기 때문이다.

Table 3.2 Association rule thresholds sorted by confidence of Park's method

player	support			confidence			lift		
	general	Lim	Park	general	Lim	Park	general	Lim	Park
SON ⇒ LEE	0.5625	0.0596	0.1489	0.8076	0.2693	0.8697	1.0279	1.0384	4.6417
KIM ⇒ HONG	0.5894	0.0639	0.1616	0.8115	0.2787	0.8682	1.0280	0.9463	4.1982
LEE ⇒ HONG	0.6095	0.0719	0.1816	0.7619	0.2687	0.8608	0.9876	0.9943	4.0196
SON ⇒ HONG	0.5510	0.0596	0.1448	0.7941	0.2824	0.8562	0.9977	0.9968	4.5564
CHO ⇒ LEE	0.6074	0.0609	0.1597	0.8125	0.2370	0.8554	0.9992	0.8805	4.2420
GAR ⇒ HONG	0.4848	0.0438	0.1159	0.8275	0.2452	0.8509	1.0503	0.8898	4.8345
KANG ⇒ LEE	0.5675	0.0491	0.1316	0.8076	0.2324	0.8500	0.9961	0.8925	4.8045
KIM ⇒ LEE	0.5495	0.0549	0.1444	0.8026	0.2477	0.8490	1.0124	0.9570	4.5568
HONG ⇒ LEE	0.6095	0.0719	0.1816	0.7901	0.2661	0.8481	0.9876	0.9943	4.0196
KANG ⇒ CHO	0.5319	0.0512	0.1317	0.7812	0.2477	0.8383	1.0343	0.9384	4.6925
GAR ⇒ CHO	0.4226	0.0440	0.1138	0.7321	0.2486	0.8202	0.9863	0.9527	4.7837
GAR ⇒ LEE	0.4521	0.0409	0.1083	0.7647	0.2244	0.8095	0.9456	0.8495	4.7595
CHO ⇒ HONG	0.5652	0.0635	0.1588	0.7647	0.2422	0.8036	0.9771	0.8820	3.9269
GAR ⇒ KANG	0.4380	0.0388	0.1030	0.7419	0.2094	0.8022	1.0671	1.0051	5.5144
KANG ⇒ HONG	0.5368	0.0481	0.1294	0.7727	0.2269	0.8006	1.0195	0.8958	4.7174
LEE ⇒ SON	0.5625	0.0596	0.1489	0.7159	0.2301	0.7947	1.0279	1.0384	4.6417
LEE ⇒ CHO	0.6074	0.0609	0.1597	0.7471	0.2263	0.7920	0.9992	0.8805	4.2420
SON ⇒ CHO	0.4742	0.0490	0.1122	0.7187	0.2364	0.7904	0.9819	0.9557	4.9514
KANG ⇒ SON	0.4952	0.0471	0.1118	0.7222	0.2267	0.7895	0.9848	0.9443	4.9663
KIM ⇒ CHO	0.5368	0.0509	0.1299	0.7500	0.2138	0.7881	1.0178	0.8477	4.5372

다음으로는 본 논문에서 제안한 신뢰도의 크기순으로 상위 20개의 규칙을 정렬한 결과를 Table 3.2에 제시하였다. 여기서 신뢰도는 고려하는 두 선수 중 한 선수가 안타를 치는 경우에 다른 선수도 안타를 치는 비율을 의미한다. 기존의 연관성 규칙에서는 SON ⇒ LEE의 규칙과 KANG ⇒ LEE의 규칙의 신뢰도가 0.8076인 반면에 Lim 등의 연구에서의 신뢰도는 각각 0.2693과 0.2324로 나타났으며, 본 논문에서 제안한 신뢰도는 각각 0.8697과 0.8500으로 계산되었다. 이로부터 알 수 있는 사실은 Lim 등의 연구 결과와 본 논문에서 제안한 신뢰도가 기존의 연관성 규칙에서의 신뢰도에 비해 좀 더 비교 가능한 정보를 제공하고 있다는 것이다. 또한 Lim 등의 연구 결과에 비해 본 논문에서 제안한 신뢰도가 기존의 신뢰도의 값과 더 비슷한 값을 가지고 있으며, 각 방법에 의한 신뢰도의 변화를 살펴보면 본 논문에서 제시한 신뢰도의 범위 (최소:0.5741, 최대:0.8509)가 가장 크므로 Lim 등의 연구 결과 (최소:0.1476, 최대:0.2787)에 비해 본 논문에서 제안한 연관성 규칙 평가 기준이 좀 더 비교 가능한 정보를 제공한다고 할 수 있다. 또한 위의 표에서 보는 바와 같이 본 논문에서 제안한 신뢰도가 Lim 등의 신뢰도에 비해 기존의 신뢰도와 더 비슷한 값으로 나타나고 있다. 그 이유는 본 논문에서 제안한 신뢰도의 공식은 기존의 신뢰도와 유사하나, Lim 등의 신뢰도는 발생항목수로 한 번 더 나누어줌으로써 기존의 신뢰도보다 작은 값으로 나타나기 때문이다.

마지막으로 각 방법에 의한 향상도의 변화를 살펴보면 기존의 연관성 규칙에서는 향상도의 최소값과 최대값이 각 0.9107과 1.0671이므로 그 범위가 0.1564로 계산되었으며, Lim 등의 연구에서의 최소 향상도와 최대 향상도는 각각 0.7849와 1.0384이므로 범위가 0.2535이다. 또한 본 논문에서 제시한 향상도의 최소값과 최대값은 각각 3.9269과 5.5144로 그 범위는 1.5875로 나타났다. 이로부터 알 수 있는 바와 같이 본 논문에서 제안한 향상도의 범위가 기존의 연관성 규칙에서의 향상도나 Lim 등의 연구에서의 향상도의 범위보다 크게 나타나고 있어서 본 논문에서 제시한 향상도가 좀 더 비교 가능한 정보를 제공하는 동시에 향상도의 비교를 용이하게 한다고 할 수 있다. 여기서 향상도는 고려하는 A, B 두 선수 중 A가 안타를 친 경우에 B도 안타를 치는 경우와 B가 임의로 안타를 치는 경우의 비를 의미한다. 한편, 본 논문에서 제안하는 향상도 보다 Lim 등의 연구 결과가 기존의 향상도와 더 비슷

한 값을 갖고 있는 것으로 나타났다. 그 이유는 Lim 등의 연구에서의 향상도의 계산 공식은 기존의 향상도의 공식과 유사한 반면에 본 논문에서 제안한 향상도는 기존의 향상도 공식에 발생항목수를 한 번 더 반영함으로써 기존 향상도보다 큰 값으로 나타나기 때문이다.

4. 결론

방대한 양의 데이터베이스로부터 쉽게 드러나지 않는 의미 있는 정보를 생성하는 데이터 마이닝 기법 중에서 기존의 연관성 규칙은 발생 여부를 나타내는 자료를 이용하여 지지도, 신뢰도, 향상도 등의 흥미도 측도를 이용함으로써 항목들 간의 관련성을 나타낸다. 이러한 연관성 규칙은 발생 빈도의 크기를 고려하지 않음으로써 정보 손실에 의한 오류를 범할 수 있다. 이를 보완한 것으로 Lim 등 (2010)의 연구가 있으나 지지도와 신뢰도의 값이 일반적인 평가 기준에 비해 상당히 작은 값으로 나타나고 있어서 기존의 방법에 익숙한 사용자들에게는 해석 시 어려움을 느끼게 할 수 있다.

이를 위해 본 논문에서는 발생 가능한 규칙의 수를 고려한 연관성 평가 기준들을 제안하고 실제 예제를 통하여 기존 연구와 비교한 후, 본 논문에서 제안한 연관성 평가 기준의 유용성을 살펴보았다. 그 결과, Lim 등의 연구 결과와 본 논문에서 제안하는 평가 기준은 기존의 연관성 규칙 평가 기준에 비해 좀 더 비교 가능한 정보를 제공하는 것을 알 수 있었다. 또한 기존의 연관성 규칙 평가 기준과 Lim 등의 연구 결과는 관심항목 수와 트랜잭션의 수를 2배로 하여도 지지도와 신뢰도, 향상도의 값은 동일한 반면에 본 논문에서 제안한 평가 기준은 발생 가능한 규칙의 수를 고려하기 때문에 트랜잭션의 수에 따라 각각의 평가 기준의 값들이 변한다는 것을 알 수 있었다. 따라서 본 논문에서 제안한 각 평가 기준이 기존의 연관성 규칙은 물론이고 Lim 등의 연구 결과에 비해서도 좀 더 비교 가능한 정보를 제공하게 된다. 연관성 평가에서 가장 중심적인 측도가 신뢰도인데, Lim 등의 연구 결과에 비해 본 논문에서 제안한 신뢰도가 기존의 신뢰도의 값과 더 비슷한 값을 가진다는 사실을 확인할 수 있었다. 각 방법에 의한 신뢰도의 변화를 살펴보면 본 논문에서 제시한 신뢰도의 범위가 가장 크므로 Lim 등의 연구 결과에 비해 본 논문에서 제안한 연관성 규칙 평가 기준이 좀 더 비교 가능한 정보를 제공한다고 할 수 있다. 끝으로, 본 논문에서 제시한 향상도의 범위가 기존의 연관성 규칙에서의 향상도나 Lim 등의 연구에서의 향상도의 범위보다 크게 나타나고 있어서 본 논문에서 제시한 향상도가 좀 더 비교 가능한 정보를 제공하는 동시에 향상도의 비교를 용이하게 한다고 할 수 있다.

참고문헌

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.

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. (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.

Lim, J., Lee, K. and Cho, Y. (2010). A study of association rule by considering the frequency, *Journal of the Korean Data & Information Science Society*, **21**, 1061-1069.

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. (2010a). Weighted association rules considering item RFM scores. *Journal of the Korean Data & Information Science Society*, **21**, 1147-1154.

Park, H. C. (2010b). Standardization for basic association measures in association rule mining. *Journal of the Korean Data & Information Science Society*, **21**, 891-899.

Park, H. C. (2010c). Decision process for right association rule generation. *Journal of the Korean Data & Information Science Society*, **21**, 263-270.

Park, H. C. (2011a). The application for predictive similarity measures of binary data in association rule mining. *Journal of the Korean Data & Information Science Society*, **22**, 495-503.

Park, H. C. (2011b). The application of some similarity measures to association rule thresholds. *Journal of the Korean Data Analysis Society*, **13**, 1331-1342.

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.

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

Wu, X., Zhang, C. and Zhang, S. (2004). Efficient mining of both positive and negative association rules. *ACM Transactions on Information Systems*, **22**, 381-405.

Journal of the Korean Data & Information Science Society
2012, **23**(4), 717–725

http://dx.doi.org/10.7465/jkdi.2012.23.4.717
한국데이터정보과학회지

Association rule thresholds considering the number of possible rules of interest items

Hee-Chang Park[1]

[1]Department of Statistics, Changwon National University

Received 12 June 2012, revised 26 June 2012, accepted 8 July 2012

Abstract

Data mining is a method to find useful information for large amounts of data in database. One of the well-studied problems in data mining is exploration for association rules. Association rule mining searches for interesting relationships among items in a given database by support, confidence, and lift. If we use the existing association rules, we can commit some errors by information loss not to consider the size of occurrence frequency. In this paper, we proposed a new association rule thresholds considering the number of possible rules of interest items and compare with existing association rule thresholds by example and real data. As the results, the new association rule thresholds were more useful than existing thresholds.

Keywords: Association rule, confidence, lift, number of possible rules, support.

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