Journal of the Korean Data & Information Science Society
2010, **21**(6), 1061–1069

한국데이터정보과학회지

발생빈도를 고려한 연관성분석 연구[†]

임제순[1] · 이경준[2] · 조영석[3]

[123]부산대학교 통계학과

접수 2010년 8월 16일, 수정 2010년 10월 22일, 게재확정 2010년 10월 27일

요 약

데이터마이닝 분야에서 연관성분석은 가장 많이 사용되고 있는 기법으로 데이터 내에 포함되어 있는 특정 항목들의 연관성을 수치화시켜 나타내는 방법이다. 기본적으로 연관성규칙은 지지도, 신뢰도, 향상도를 계산하여 연관성의 유무를 판단한다. 기존에 제시된 관련 논문에서는 관심변수의 발생유무만을 바탕으로 연관성규칙을 이용하였고, 빈번하지 않은 데이터에 대한 문제점과 순위결정함수를 통한 해결방안을 제시하였다. 하지만 실제 데이터에서는 발생이 빈번하지 않은 데이터 뿐 아니라, 발생이 많이 일어나는 데이터도 존재한다. 따라서 발생빈도를 고려한 연관성규칙이 필요하다고 생각한다.본 논문에서는 각 케이스 내의 발생빈도를 고려한 새로운 연관성 측정 도구를 제시하였다. 또한 실제 예제를 통하여 기존의 연관성규칙과 새로운 연관성규칙의 결과를 비교해 보았다. 그 결과, 새로 제시한 연관성규칙이 기존의 연관성규칙보다 더 세밀하게 구분하는 것을 확인할 수 있었다.

주요용어: 데이터마이닝, 발생빈도, 연관성, 연관성규칙.

1. 서론

현대 사회에서 정보 (information)는 매우 중요한 역할을 하고 있으며, 정보획득을 위해 다양한 통계적 방법론이 연구되어왔다. 그 중 데이터마이닝 (data mining)은 대규모 데이터를 바탕으로 그 안에 포함되어 있는 유용한 정보를 얻는 과정이며, 데이터로부터 특정한 연관성을 발견하는 것은 데이터마이닝의 가장 일반적인 작업이라 할 수 있다. 연관성규칙은 하나의 거래나 사건에 포함되어 있는 항목들 간의 관련성을 파악하여 둘 이상의 항목들로 구성된 연관성규칙을 도출하는 탐색적 자료분석의 방법으로, Agrawal 등 (1993)에 의해 가장 먼저 소개되었으며, 이후 많은 학자들이 이와 관련된 연구를 수행하였다 (Agrawal과 Srikant, 1994; Wu와 Zhang, 2004; Cho와 Park, 2007; Park, 2010a; Park, 2010b; Park, 2010c; Park, 2010d).

일반적으로 연관성규칙을 이해하기는 어려운 점이 없으나 대용량의 데이터로부터 얻어지는 여러 가지 연관성규칙들은 모두 의미 있는 내용을 포함하고 있지는 않다. 타당한 근거 없이 우연히 발생되는 경우에는 일반화에 무리가 있다. 이러한 경우 의미 있는 연관성을 찾아내기 위해서는 각각의 연관성을 비교할 수 있는 규칙이 필요하다. 연관성규칙은 기본적으로 지지도 (support), 신뢰도 (confidence), 향상도 (lift)를 평가도구로 사용하여 판단하게 된다. 각각의 경우에 따라 지지도, 신뢰도, 향상도를 구하고 임의의 최소 지지도, 최소 신뢰도, 향상도를 기준으로 연관성의 유무에 대해 평가한다.

[†] 이 논문은 부산대학교 자유과제 학술연구비 (2년)에 의하여 연구되었음.
[1] (609-735) 부산광역시 금정구 부산대학로 63번길 2, 부산대학교 통계학과, 석사과정.
[2] (609-735) 부산광역시 금정구 부산대학로 63번길 2, 부산대학교 통계학과, 박사과정.
[3] 교신저자: (609-735) 부산광역시 금정구 부산대학로 63번길 2, 부산대학교 통계학과, 부교수.
E-mail: choys@pusan.ac.kr

하지만 실제 데이터의 특성을 좀 더 감안한다면 우리는 새로운 문제점에 봉착하게 된다. 연관성분석은 관심변수의 발생 유무를 바탕으로 지지도, 신뢰도, 향상도를 구하고 있다. 즉, 각 케이스별로 해당 관심변수의 발생 유무를 기초로 한 지지도, 신뢰도, 향상도를 얻고 사용해왔다. 그러나 실제 데이터에서는 케이스 내에서 관심변수의 발생빈도가 빈번하지 않은 경우도 있고, 발생빈도가 많은 경우도 있다. 발생빈도가 적은 경우에 대해서는 순위결정함수 (Wu와 Zhang, 2004; Park, 2010a; Park, 2010b)를 통해 해결방안을 제시해왔지만 발생빈도가 많은 경우에는 아직 언급된 바가 없다.

실제 데이터를 분석하다보면 각 케이스별로 관심변수가 하나 이상의 양적인 의미를 내포하고 있는 경우가 많은데, 연관성규칙에서 가장 빈번하게 활용되는 마케팅 분야에서 살펴보더라도 그 현상은 쉽게 나타난다. 소액 품목의 경우 1회 거래 내에서 구매품목당 구매 개수가 2 이상인 경우가 많이 나타나고 있지만 연관성분석 과정에서는 단지 구매 유무만을 가지고 판단하고 있다.

이런 과정을 통해 연관성분석에서 의도치 않은 데이터의 손실이 일어나게 되고, 잘못된 판단 또는 세밀한 구분을 하지 못하는 경우가 생기게 된다. 이러한 문제점을 해결하기 위해 본 논문에서는 발생빈도를 고려한 새로운 연관성규칙을 제시하였다.

본 논문 2장에서는 기존의 연관성규칙에 대한 설명과 예제를 통한 문제점에 대해 살펴보고, 3장에서는 기존의 문제점을 보완하는 새로운 연관성규칙을 제시하였다. 그리고 4장에서는 실제 예제를 통해 기존의 연관성분석 방법과 새로운 방법의 결과를 각각 비교해 보며, 5장에서는 결론 및 앞으로의 연구방향을 제시하겠다.

2. 기존의 연관성규칙과 문제점

연관성규칙을 판단하는 기준은 기본적으로 지지도, 신뢰도, 향상도 세 가지로 구성되어 있다. 모든 연관성규칙의 기초가 되는 지지도는 전체에 대한 항목 A와 항목 B가 동시에 일어나는 확률을 의미하고 식으로 나타내면 아래와 같다 (강현철 등, 2006).

$$\text{Support}(A \Rightarrow B) = \Pr(A \cap B) = \frac{\text{항목A와B를동시에포함하는케이스수}}{\text{전체케이스수}}. \tag{2.1}$$

지지도의 경우 Support(A⇒B)와 Support(B⇒A)가 상호 대칭적으로 서로 같은 값을 가진다. 즉, A와 B는 서로 연관 상대인 다른 항목에 대한 비중에 영향을 받게 된다. 이러한 이유로 지지도는 관심변수 모두의 비중이 크고, 연관성도 큰 경우에는 유용하게 사용되지만, 관심변수의 전체에 대한 포함비중이 낮은 경우에는 연관성을 판단하는데 어려움이 있다. 이러한 지지도의 단점을 보완하는 것이 신뢰도이며, 신뢰도는 A가 발생한 경우 중 B가 발생하는 경우의 조건부확률을 의미하고 식으로 나타내면 아래와 같다.

$$\begin{aligned}\text{Confidence}(A \Rightarrow B) = \Pr(B|A) &= \frac{\Pr(A \cap B)}{\Pr(A)} \\ &= \frac{\text{항목 A와 B를 동시에 포함하는 케이스 수}}{\text{항목 A를 포함하는 케이스 수}}.\end{aligned} \tag{2.2}$$

위의 식에서 보면 알 수 있듯이 신뢰도는 지지도와는 달리 대칭적이지 않다. 지지도 또는 신뢰도가 높은 연관성규칙 중에는 우연하게 연관성이 높게 보이는 것들이 나타날 수도 있는데, 이 부분을 보완하기

위해서 향상도가 사용된다. 향상도는 식으로 나타내면 아래와 같다.

$$\begin{aligned}\mathrm{Lift}(A \Rightarrow B) &= \frac{\Pr(B|A)}{\Pr(B)} = \frac{\Pr(A \cap B)}{\Pr(A)\Pr(B)} \\ &= \frac{\text{항목 A와 B를 동시에 포함하는 케이스 수} \times \text{전체 케이스 수}}{\text{항목 A를 포함하는 케이스 수} \times \text{항목 B를 포함하는 케이스 수}}. \end{aligned} \tag{2.3}$$

위의 세 가지 수식에서 나타난 바와 같이 지지도, 신뢰도, 향상도는 서로 밀접한 관계를 가지고 있으며, 그 관계성은 쉽게 이해할 수 있다. 또한 앞서 서론에서 말한 바와 같이 기존의 지지도, 신뢰도, 향상도는 어떤 특정한 사건이 일어났는지의 여부만을 기준으로 하여 계산되고 있음을 알 수 있다. 즉, 특정한 케이스 내에서 특정한 사건이 몇 번 일어났느냐에 관한 측면에서는 데이터의 손실이 있다고 할 수 있다. 이 문제는 아래의 표 2.1의 예제를 보면 쉽게 이해할 수 있다.

표 2.1 예제 1

케이스 (날짜)	데이터 1	데이터 2
1	**AB**	**AB** CCCCDEF
2	**AB**	**AAAB**
3	**AB**	**ABBB**
4	**AB**	**AAABBB**
5	**AB**	**ABAB** C

케이스가 하나인 경우에 대해 생각해보자. 각 데이터의 케이스 1만을 대상으로 생각했을 때, 데이터 1의 경우 Support(A⇒B)=1이고, 데이터 2의 경우에도 Support(A⇒B)=1이다. 지지도뿐만 아니라 신뢰도, 향상도의 경우에도 두 경우 모두 동일한 값을 가진다. 그러나 각각 케이스 내에서 관심변수가 차지하는 비중을 고려해 볼 때, 두 데이터의 연관성은 다르다고 생각하는 것이 명백하다.

이와 같은 문제점은 실제의 데이터로 생각해본다면 더욱 쉽게 이해할 수 있다. 표 2.1을 프로야구 선수들의 안타를 기록한 데이터라고 생각해보자. 데이터 1의 경우 A선수와 B선수는 매 경기 (각 케이스) 각각 안타를 1개씩 쳤으며, 매 경기 그 팀에서 안타는 단 두 개씩, 즉 A, B 선수가 하나씩 안타를 친 것이 전부이다.

데이터 2의 경우에는 매 경기마다 다른 결과를 가지고 있다. 케이스 1의 경우 데이터 1과 같이 A선수와 B선수는 각각 안타를 하나씩 쳤지만, 다른 선수들도 그날 안타를 쳤다. 앞에서 언급하였듯이 케이스 내에서 관심변수 (A, B 선수의 안타)가 차지하는 비중은 다르게 느껴지므로 지지도에 문제가 있다고 생각된다. 케이스 2와 케이스 3의 경우에는 지지도의 측면에서는 문제가 없다고 생각되나 신뢰도를 생각해보면, A와 B 쌍방의 신뢰도는 다르게 생각될 수 있으나 신뢰도 값은 같음을 알 수 있다. 케이스 4의 경우에는 관심변수의 전체에 대한 비중은 같게 생각되지만 연관성의 개념 자체만을 생각해보면 데이터 1보다 데이터 2가 더 연관성이 높다고 생각된다. 케이스 5의 경우에는 앞에서 언급한 모든 문제점이 포함되어있는 경우이다.

위의 데이터를 바탕으로 각각 지지도, 신뢰도, 향상도를 계산해보면 데이터 1과 데이터 2 모두 같은 연관성 값을 가짐을 알 수 있다. 데이터 1의 경우 Support(A⇒B)=5/5=1, Confidence(A⇒B)=1/1=1, Lift(A⇒B)=1/1=1이고, 데이터 2의 경우에도 마찬가지로 Support(A⇒B)=5/5=1, Confidence(A⇒B)=1/1=1, Lift(A⇒B)=1/1=1임을 구할 수 있다. 즉, 데이터 1과 데이터 2가 상황이 다르더라도 현재 연관성규칙을 판단하는 척도로 사용되는 지지도, 신뢰도, 향상도는 같은 값을 가진다.

이러한 문제점은 연관성규칙을 이용하는 모든 경우 빈번히 나타날 수 있는 문제점이므로 해당 케이스 내에서의 관심 변수에 대한 발생빈도를 고려하는 새로운 방법이 필요하다고 생각된다.

3. 새로운 연관성규칙 제시

데이터마이닝에서 연관성규칙을 발생빈도가 작은 실제 데이터에 적용하면 모호한 부분이 많이 존재한다. 그 방법을 보완하기 위해서 여러 가지 순위결정함수들이 제시되어왔다. 하지만 앞서 제시한 바와 같이 발생빈도가 많은 실제 데이터의 문제점들은 순위결정함수를 통해서 해결되지 않는다. 이 역시 기존의 연관성규칙을 이용해 케이스 내의 사건의 비중에 대해서는 무시하고 사건의 발생 유무만을 판단한 것을 사용하고 있다. 이 방법을 보완하기 위해 새로운 연관성규칙을 제시한다.

먼저 지지도는 연관성규칙에서 가장 기초가 되는 측정도구로 기존의 경우 분자는 A와 B를 모두 포함한 케이스의 개수가 들어가게 되고, 유효한 케이스 당 1의 값을 가지게 된다. 여기서 한 케이스 내의 관심변수의 발생빈도를 고려한 가중치를 부여하여 기존의 문제를 해결하는 방법을 생각해 보았다. 수식 (2.1)에서 분자를 변형하여 다음과 같은 새로운 지지도 (Support_m)를 제시한다. 여기서 명명한 아래첨자의 'm'은 수정되었다는 의미의 단어인 'modified'의 첫 글자를 이용하여 표기하였다.

$$\text{Support}_m(A \Rightarrow B) = \frac{\sum_{case}\left(\frac{\#A}{\#Total}\times 2\times\frac{\#B}{\#Total}\times 2\right)}{\text{전체 케이스 수}}, \tag{3.1}$$

여기서 #은 케이스 당 해당 값의 수를 나타낸다. 즉, # Total은 케이스 내에 일어난 모든 값의 개수가 되고, # A는 케이스 내에 A의 개수, # B는 케이스 내에 B의 개수를 의미한다. 제시한 새로운 지지도의 계산법은 케이스 내의 관심변수의 발생빈도 누락 문제점을 해결할 수 있다. 또한 기존의 연관성규칙에서는 최소 지지도, 최소 신뢰도, 향상도에 따른 의사결정만을 할 수 있었는데, 제시한 방법을 통해서는 의사결정 뿐만 아니라 각 경우에 따라 연관성규칙의 정도를 측정하는 측도로 사용될 수 있을 것으로 예상되며, 어느 경우가 연관성이 높은지 비교도 가능하다. (3.1) 식에서 각각의 비중에 2를 곱해주는 것은 앞의 표 2.1의 데이터 1과 같이 순수 관심변수로 구성된 경우에 지지도 값을 1로 맞춰주기 위해서이다. 사건 A와 사건 B에 대한 지지도의 경우 (3.1)의 식과 같으나 사건을 여러 개로 확장한 경우 일반적으로 정리하면 아래와 같다.

$$\text{Support}_m(A_1 \Rightarrow \cdots \Rightarrow A_k) = \frac{\sum_{case}\left\{\prod_{i=1}^{k}\left(\frac{A_i}{\#Total}\times k\right)\right\}}{\text{전체 케이스 수}}. \tag{3.2}$$

기존의 신뢰도 계산방법 역시 한 케이스 내에 차지하는 관심변수의 발생빈도를 고려하지 않았으므로 문제점이 발생할 수 있다. 표 2.1의 데이터 2에서 케이스 2와 케이스 3에 대해서 생각해보자. 케이스 2의 경우 (AAAB)로 구성되어 있으며, 케이스 3의 경우 (ABBB)로 구성되어 있다. 각각 기존의 신뢰도인 P(B|A)를 계산해보면, 두 경우 모두 Confidence(A⇒B)=1로 동일한 값을 가지게 되며, 데이터를 5개로 확장한 표 3.1의 예제에서도 Confidence(A⇒B)의 값은 각각 1로 동일하다.

표 3.1 예제 2

데이터 1	데이터 2
AAAB	ABBB
AAAB	ABBB
AAAB	ABBB
AAAB	ABBB
AAAB	ABBB

앞서 지지도에서 접근했던 방식과 마찬가지로 케이스 내의 관심변수의 발생빈도를 고려하여 새로운 신뢰도 (Confidence$_m$)를 제시하면 아래 수식과 같다.

$$\begin{aligned}\text{Confidence}_{\text{m}}(A \Rightarrow B) &= \frac{\sum_{case}\left(\frac{\#A}{\#Total}\times 2\times\frac{\#B}{\#Total}\times 2\right)}{\sum_{case}\left(\frac{\#A}{\#Total}\times 2\right)} \\ &= \frac{\sum_{case}\left(\frac{\#A}{\#Total}\times\frac{\#B}{\#Total}\times 4\right)}{\sum_{case}\left(\frac{\#A}{\#Total}\times 2\right)}.\end{aligned} \tag{3.3}$$

새로 제안한 Confidence$_m$ (A$\Rightarrow$ B)를 이용하여 표 3.1의 데이터를 적용시켜 보면, 데이터 1의 경우에는 Confidence$_m$ (A$\Rightarrow$B)=(3/4)/(3/2)=1/2, 데이터 2의 경우에는 Confidence$_m$ (A$\Rightarrow$B)=(3/4)/(1/2)=3/2으로 데이터 2의 경우가 신뢰도 측면에서 더 높게 나타남을 알 수 있다. 즉, 새로 제안한 신뢰도는 기존 신뢰도의 문제점을 해결할 수 있는 것으로 나타났다.

앞의 (3.1), (3.2)를 바탕으로 기존의 향상도에서 케이스 내의 관심변수의 발생빈도를 고려하여 새로운 향상도 (Lift$_m$)는 아래와 같다.

$$\text{Lift}_{\text{m}}(A \Rightarrow B) = \frac{\sum_{case}\left(\frac{\#A}{\#Total}\times\frac{\#B}{\#Total}\times 4\right)\times \text{전체 케이스 수}}{\sum_{case}\left(\frac{\#A}{\#Total}\times 2\right)\sum_{case}\left(\frac{\#B}{\#Total}\times 2\right)}. \tag{3.4}$$

새로 제시한 지지도, 신뢰도, 향상도를 이용하여 표 2.1의 데이터를 분석하고, 기존의 방식과 비교한 결과 아래 표 3.2와 같다.

표 3.2 기존의 연관성과 새로운 연관성 비교 (예제 1)

구분		데이터 1	데이터 2
지지도	Support	1	1
	Support$_m$	1	0.637877
신뢰도	Confidence	1	1
	Confidence$_m$	1	0.79294
향상도	Lift	1	1
	Lift$_m$	1	0.985699

위의 표 3.2 결과에서 볼 수 있듯이 기존의 연관성 규칙보다는 새로 제시한 연관성 규칙이 데이터의 상황을 보다 더 세밀하게 반영하고 있는 것을 알 수 있다.

4. 예제를 통한 고찰

예제 데이터는 2010 한국프로야구 롯데 자이언츠 1군 선수들의 상반기 89경기의 매 경기 안타의 데이터를 이용하였다 (출처: www.giantsclub.com).

아래 표 4.1은 기존의 지지도, 신뢰도, 향상도를 구하여, 향상도가 1 이상인 규칙 (Rule) 중 지지도, 신뢰도가 높은 순서로 정렬하였다 (상위 30개). 단 포함률이 5% 이하이거나, 신뢰도가 10% 이하인 경우에는 제외하였다.

표 4.1 2010시즌 상반기 롯데자이언츠 타자 데이터 연관성 분석

순위	규칙	Support	Confidence	Lift
1	손아섭==>홍성흔	55.56	79.37	1.01
2	홍성흔==>손아섭	55.56	70.42	1.01
3	손아섭==>이대호	52.22	74.6	1
4	이대호==>손아섭	52.22	70.15	1
5	김주찬==>홍성흔	47.78	82.69	1.05
6	홍성흔==>김주찬	47.78	60.56	1.05
7	조성환==>이대호	45.56	78.85	1.06
8	강민호==>이대호	45.56	75.93	1.02
9	이대호==>조성환	45.56	61.19	1.06
10	이대호==>강민호	45.56	61.19	1.02
11	가르시아==>홍성흔	44.44	81.63	1.03
12	홍성흔==>가르시아	44.44	56.34	1.03
13	강민호==>손아섭	43.33	72.22	1.03
14	손아섭==>강민호	43.33	61.9	1.03
15	박종윤==>홍성흔	36.67	82.5	1.05
16	홍성흔==>박종윤	36.67	46.48	1.05
17	박종윤==>이대호	34.44	77.5	1.04
18	박종윤==>가르시아	34.44	77.5	1.42
19	가르시아==>박종윤	34.44	63.27	1.42
20	가르시아==>강민호	34.44	63.27	1.05
21	강민호==>가르시아	34.44	57.41	1.05
22	이대호==>박종윤	34.44	46.27	1.04
23	전준우==>홍성흔	32.22	80.56	1.02
24	전준우==>이대호	32.22	80.56	1.08
25	이대호==>전준우	32.22	43.28	1.08
26	홍성흔==>전준우	32.22	40.85	1.02
27	박종윤==>강민호	30	67.5	1.13
28	강민호==>박종윤	30	50	1.13
29	박기혁==>이대호	26.67	92.31	1.24
30	전준우==>강민호	26.67	66.67	1.11

아래 표 4.2는 새로 제안한 지지도 (Support_m), 신뢰도 (Confidence_m), 향상도 (Lift_m)를 구하여, 지지도 (Support_m), 향상도 (Confidence_m)가 높은 순서로 정렬하였다 (상위 30개). 단, 적용된 데이터의 경우 각 관심변수의 빈도가 각 케이스 내에 포함되는 비중이 크지 않고, 그 두 값의 곱을 통해 계산되기 때문에 제안한 지지도 (Support_m), 신뢰도 (Confidence_m)를 계산하게 되면 그 값이 매우 작게 나타나게 된다. 비교의 용의성을 위해 지지도 (Support_m), 신뢰도 (Confidence_m)에 각각 100을 곱하여 표에 수록하였다.

표 4.1과 표 4.2의 결과로부터, 기존 지지도, 신뢰도, 향상도보다 새로 제시한 지지도, 신뢰도, 향상도가 좀 더 세밀하고 정확하게 나타남을 알 수 있다. 특히 기존의 지지도의 경우에는 데이터의 케이스 수가 적으면, 위에서 보는 바와 같이 동일한 값을 가지는 규칙이 여러 개 생겨서, 분별할 수 없는 경우가 많이 나타난다 (순위 7∼10, 17∼22, 23∼26). 그러나 새로 제시한 지지도의 경우에는 짝을 이루는 값을 제외하고 같은 값을 가지는 규칙은 존재하지 않는다. 또한 기존의 연관성 분석을 통해 계산된 값의 순위가 뒤바뀌는 규칙도 생긴다. 예를 들어 표 4.1에서 순위 3, 4와 순위 5, 6의 규칙의 경우 새로 제안한 연관성규칙을 이용하면 서로 순위가 바뀌는 것을 볼 수 있다. 새로 제시한 연관성규칙이 케이스 내의 발생

표 4.2 2010시즌 상반기 롯데자이언츠 타자 데이터에 대한 새로운 연관성 분석

순위	규칙	Support_m	Confidence_m	Lift_m
1	손아섭==>홍성흔	6.38	28.80	1.02
2	홍성흔==>손아섭	6.38	22.71	1.02
3	김주찬==>홍성흔	5.87	29.79	1.06
4	홍성흔==>김주찬	5.87	20.88	1.06
5	손아섭==>이대호	5.64	25.43	0.98
6	이대호==>손아섭	5.64	21.71	0.98
7	조성환==>이대호	5.07	23.93	0.92
8	이대호==>조성환	5.07	19.55	0.92
9	강민호==>이대호	4.38	22.69	0.87
10	이대호==>강민호	4.38	16.88	0.87
11	가르시아==>홍성흔	4.29	24.49	0.87
12	홍성흔==>가르시아	4.29	15.26	0.87
13	강민호==>손아섭	4.12	21.31	0.96
14	손아섭==>강민호	4.12	18.57	0.96
15	강민호==>가르시아	3.22	18.41	0.95
16	전준우==>홍성흔	3.22	16.70	0.95
17	홍성흔==>박종윤	3.17	25.03	0.89
18	이대호==>전준우	3.17	11.27	0.89
19	이대호==>박종윤	3.05	24.12	1.38
20	전준우==>이대호	3.05	17.43	1.38
21	박종윤==>이대호	3.02	27.30	1.05
22	가르시아==>강민호	3.02	11.64	1.05
23	박종윤==>가르시아	2.84	22.45	0.86
24	박종윤==>홍성흔	2.84	10.94	0.86
25	가르시아==>박종윤	2.47	22.30	0.79
26	홍성흔==>전준우	2.47	8.78	0.79
27	박기혁==>이대호	2.29	20.72	1.05
28	강민호==>박종윤	2.29	11.64	1.05
29	전준우==>김주찬	2.27	17.90	0.93
30	전준우==>강민호	2.27	11.73	0.93

빈도를 고려하여 계산하였으므로 기존의 연관성규칙보다 더 세밀하게 데이터의 상황을 반영하여 연관성을 나타낸다고 할 수 있다. 단, 기존의 연관성 분석에서 나타나는 향상도는 1보다 큰 것을 기준으로 사용하는데 새로 제안한 향상도의 경우에는 기준값의 필요성에 대해서는 앞으로 연구가 필요하다고 생각된다.

표 4.3 기존의 연관성규칙과 새로운 연관성규칙 결과

순위	Support	Support_m	Confidence	Confidence_m
1	손아섭=>홍성흔	손아섭=>홍성흔	김주찬=>홍성흔	김주찬=>홍성흔
2	홍성흔=>손아섭	홍성흔=>손아섭	손아섭=>홍성흔	손아섭=>홍성흔
3	손아섭=>이대호	김주찬=>홍성흔	전준우=>이대호	박종윤=>이대호
4	이대호=>손아섭	홍성흔=>김주찬	박기혁=>이대호	박종윤=>강민호
5	김주찬=>홍성흔	손아섭=>이대호	손아섭=>이대호	손아섭=>이대호
6	홍성흔=>김주찬	이대호=>손아섭	박종윤=>홍성흔	홍성흔=>박종윤
7	조성환=>이대호	조성환=>이대호	가르시아=>홍성흔	가르시아=>홍성흔
8	강민호=>이대호	이대호=>조성환	박종윤=>가르시아	이대호=>박종윤
9	이대호=>조성환	강민호=>이대호	조성환=>이대호	조성환=>이대호
10	이대호=>강민호	이대호=>강민호	박기혁=>손아섭	박기혁=>손아섭

표 4.3은 표 4.1과 표 4.2를 바탕으로 하여 기존 지지도 (support)와 새로 제시한 지지도 (Support_m),

기존 신뢰도 (confidence)와 새로 제시한 신뢰도 (Confidence$_m$)의 상위 10개의 규칙을 정리하여 비교해 보았다.

지지도의 측면에서 보면 '김주찬=>홍성흔'이 5위에서 3위로 상승하였고, '홍성흔=>김주찬'이 6위에서 4위로 상승하였으며, '이대호=>조성환'이 9위에서 8위로 상승한 것을 볼 수 있다. 신뢰도의 측면에서 보더라도 '박종윤=>이대호', '박종윤=>강민호', '홍성흔=>박종윤', '이대호=>박종윤'이 10위권 밖에서 10위권 내의 순위로 상승한 것을 알 수 있다.

5. 결론

데이터마이닝에서 최소 지지도, 최소 신뢰도를 만족하며, 향상도가 1 이상인 경우에 보통 연관성규칙이 있다고 판단한다. 실제 데이터를 통해 데이터마이닝 연관성 분석을 하면 문제점이 발생하게 된다. 먼저 관심 변수의 빈도가 적은 경우에는 세 가지 기준을 모두 충족시키기 힘들다. 이러한 경우를 해결하는 방안으로 여러가지 순위결정함수가 제안되었다. 또한 실제 데이터의 특성상 빈도가 적은 경우는 물론 빈도가 많은 경우도 있다.

본 논문에서는 각 케이스별 관심변수의 발생빈도를 통해 수치적인 측면을 고려한 새로운 지지도, 신뢰도, 향상도를 제안하고, 또한 실제 데이터를 통한 기존 연관성분석과 새로 제안한 연관성분석 결과의 차이를 살펴보았다. 그 결과 기존의 연관성규칙을 통해서 분별할 수 없었던 부분들을 새로 제안한 연관성규칙을 통해 분별이 가능하였으며, 순위가 뒤바뀌는 경우도 발생하였다. 즉, 새로 제안한 연관성규칙이 기존의 연관성규칙보다 더 데이터의 상황을 세밀하게 반영하고 있다고 할 수 있다.

본 논문에서 제안된 연관성규칙을 이용하여 추가로 순위결정함수 등에도 적용시킨다면 더욱 더 유용한 연관성분석이 되리라 생각된다.

참고문헌

강현철, 한상태, 최종후, 이성건, 김은석, 엄익현, 김미경 (2006). <데이터마이닝 방법론>, 자유아카데미, 경기도.

Agrawal, R., Imielinski, T. 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 International Conference on VLDB*, 487-499.

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.

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, H. C. (2010a). Association rule ranking function by decreased lift influence. *Journal of the Korean data & Information science Society*, **21**, 397-405.

Park, H. C. (2010b). Development of associative rank decision function using basic association rule thresholds. *Journal of the Korean data Analysis Society*, **12**, 961-971.

Park, H. C. (2010c). Proposition of symmetric confidence considering relative size of item frequencies. *Journal of the Korean data Analysis Society*, **12**, 1463-1472.

Park, H. C. (2010d). Association rule ranking fuction using conditional probability increment ratio. *Journal of the Korean data & Information science Society*, **21**, 709-717.

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
2010, **21**(6), 1061–1069

한국데이터정보과학회지

A study of association rule by considering the frequency[†]

Jesoon Lim[1] · KyeongJun Lee[2] · Youngseuk Cho[3]

[123]Department of Statistics, Busan National University

Received 16 August 2010, revised 22 October 2010, accepted 27 October 2010

Abstract

In data mining, association rule is a popular and well researched method for discovering interesting relations between variables. There are three measures for association rule, support, confidence and lift. But there are some problem in them. They don't consider the frequency of variable in case. So, we need the new association rule which consider the frequency.In this paper, we proposed the new association rule. We compared the proposed association rule with the original association rule from example data. As a result, we knew our function was better than the original function in terms of sensitivity.

Keywords: Association, association rule, data mining, frequency.

[†] This work was supported for two years by Pusan National University Research Grant.
[1] Master student, Division of Mathematics and Statistics, Pusan National University, Busan 609-735, Korea.
[2] Doctor of philosophy student, Division of Mathematics and Statistics, Pusan National University, Busan 609-735, Korea.
[3] Corresponding author: Associate professor, Division of Mathematics and Statistics, Pusan National University, Busan 609-735, Korea. E-mail: choys@pusan.ac.kr