기본적인 연관평가기준 전부를 고려한 비선형 회귀모형에 의한 연관성 규칙 수의 결정

박희창¹

¹창원대학교 통계학과

접수 2013년 2월 4일, 수정 2013년 2월 20일, 게재확정 2013년 3월 1일

요 약

데이터 마이닝 기법들 중에서도 연관성 규칙은 가장 최근에 개발된 기법으로 대용량 데이터베이스에서 각 항목들 간의 관련성을 찾아내며, 두 항목간의 관계를 명확히 수치화함으로써 두 개 이상의 항목간의 관련성을 표시하여 주기 때문에 현장에서 직접 적용이 가능하다. 일반적으로 연관성 규칙 생성 여부를 판단할 때, 각 항목간의 연관성을 반영하는 기준인 지지도, 신뢰도, 향상도 등의 흥미도 측도를 활용하게 된다. 실제적으로 연관성 규칙의 수를 결정하기 위해서는 이들 흥미도 측도들의 평가 기준을 정하기 위해 반복적으로 조정 과정을 거쳐야 한다. 본 논문에서는 이러한 문제를 해결하기 위해 연관성 평가기준 모두를 일반적으로 많이 활용되고 있는 비선형 회귀모형에 적용하여 연관성 규칙의 수를 추정하는 방안을 강구하였다. 또한 분산팽창계수를 이용하여 다중공선성 문제를 진단하는 동시에 분산분석 결과와 수정 결정계수를 이용하여 각 모형의 기여도를 비교하여 가장 바람직한 회귀 모형을 구하였다.

주요용어: 비선형 회귀모형, 신뢰도, 연관성 평가기준, 지지도, 향상도, 흥미도 측도.

1. 서론

오늘날 세계는 지금 네 가지 핵심 트렌드인 빅 데이터, 모바일, 클라우드 컴퓨팅, 그리고 소셜 미디어 등에 의해 유례없는 변화의 시기를 맞이하고 있다. 이들 중에서 빅 데이터를 성공적으로 활용하기 위해서는 이를 가공하고 분석 및 처리하는 기술인 데이터마이닝 기법의 적용이 필수적이다. 이 기법은 방대한 양의 데이터에서 쉽게 드러나지 않는 유용한 정보를 찾아내는 과정을 의미하는 것으로, 이 기법을 활용함으로써 경쟁 상황 속에서 조직의 최적 전략이나 의사결정을 위한 의미 있는 고급정보를 확보할 수 있다. 데이터 마이닝 기법들 중에서도 연관성 규칙은 가장 최근에 개발된 기법으로 대용량 데이터베이스에서 각 항목들 간의 관련성을 찾아내며, 두 항목간의 관계를 명확히 수치화함으로써 두 개 이상의 항목간의 관련성을 표시하여 주기 때문에 현장에서 직접 적용이 가능하다. 특히 백화점이나 인터넷 비즈니스 환경 등에서의 시장바구니 분석, 개인화 추천 서비스, 로그파일의 분석, 통신업계에서의 교차 판매 분석, 그리고 제조업계에서의 불량판별기준의 설정 등 많은 분야에서 활용이 가능하다 (Park, 2013).

연관성 규칙 마이닝은 Agrawal 등 (1993)에 의해 처음 소개된 이후, 많은 학자들에 의해 연구가 수행된 바 있다. 이들 연구들을 분류하면 크게 두 가지로 나누어지는데, 이중 제약조건을 가지는 항목으로 구성된 트랜잭션 데이터베이스에서 빈발항목을 찾는 연구로는 Han과 Fu (1995, 1999), Srikant 등

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

(1997), Cai 등 (1998), Liu 등 (1999)이 있다. 그리고 연관성 규칙 생성에 대한 수행속도를 향상시키기 위한 연구에는 Agrawal과 Srikant (1994), Park 등 (1995), Bayardo (1998), Pasquier 등 (1999), Han 등 (2000), Pei 등 (2000), Toivonen (1996) 등이 있다. 또한 최근의 연관성 규칙에 대한 국내 연구로는 Cho와 Park (2011a, 2011b), Lim 등 (2010), Park (2010a, 2010b, 2011a, 2011b, 2011c) 등이 있다.

일반적으로 연관성 규칙 생성 여부를 판단할 때, 각 항목간의 연관성을 반영하는 기준인 지지도 (support), 신뢰도 (confidence), 향상도 (lift) 등의 흥미도 측도 (interestingness measure)를 활용하게 된다. 만약 이들에 대한 기준값들을 너무 작게 하면 너무 많은 연관성 규칙이 생성되며, 너무 크게 하면 생성되는 연관성 규칙의 수가 작아지게 된다 (Park, 2011a). 연관성 규칙 수의 추정 문제를 해결하기 위해 Yi 등 (2011)은 지지도와 신뢰도의 기준값에 대해 일반적으로 많이 활용되고 있는 비선형 회귀모형을 적용한 바 있다. 그러나 그들은 회귀모형의 기본 가정이나 특히 다중회귀모형에서 반드시 고려하여야 할 다중공선성 문제를 진단하지 않고 결론을 내린 바 있다. 그리고 Park (2013)은 이러한 문제를 보완하는 동시에 이들 기준값들의 값을 쌍으로 고려하여 연관성 규칙의 수를 추정하기 위해 Yi 등 (2011)이 제안한 비선형 회귀모형을 고려한 바 있다. 본 논문에서는 Yi 등 (2011)이 제안한 비선형 회귀모형의 문제점과 Park (2013)이 제안한 모형을 보완하기 위해 3가지 기본적인 연관성 평가기준 모두를 반영한 비선형 회귀모형을 고려하고자 한다, 본 논문의 2절에서는 연관성 규칙의 수를 결정하기 위해 적용할 수 있는 비선형 회귀 모형을 제안하고, 3절에서는 예제를 통하여 Yi 등 (2011) 및 Park (2013)에서 고려한 모형과의 비교를 통해 가장 바람직한 회귀 모형을 결정하는 방안에 대해 논의하며, 마지막으로 4절에서 결론을 내리고자 한다.

2. 연관성 규칙 수의 결정을 위한 비선형 회귀 모형의 제안

기존의 연관성 규칙의 평가기준인 지지도, 신뢰도, 향상도 등을 수식으로 나타내기 위해 Table 2.1과 같은 분할표를 고려한다 (Park, 2012).

Table 2.1 2×2 contingency table

		<i>B</i>		Total
		1	0	
<i>A</i>	1	n_{11}	n_{10}	n_{1+}
	0	n_{01}	n_{00}	n_{0+}
Total		n_{+1}	n_{+0}	n

여기서 지지도 $S(A \Rightarrow B)$ 는 항목 집합 A 와 항목 집합 B 가 동시에 발생하는 거래의 비율로 $S(A \Rightarrow B) = n_{11}/n$ 으로 계산된다. 지지도는 두 항목의 동시발생비율을 의미하나, 현실적으로 지지도의 적절한 평가 기준을 정하기가 어렵다. 만약 지지도의 평가 기준을 낮게 책정하면 의미 없는 연관성 규칙을 포함하여 많은 수의 연관성 규칙이 생성된다. 반면에 지지도의 평가 기준을 높게 책정하면 낮은 지지도를 가지는 중요한 패턴의 연관성 규칙은 생성되지 않는다. 따라서 지지도의 평가기준을 정하기 위해서는 조정 과정을 반복적으로 거쳐야 한다. 신뢰도 $C(A \Rightarrow B)$ 는 항목 집합 A 가 포함된 거래 비율 중 항목 집합 A 와 항목 집합 B 가 동시에 포함된 거래의 비율을 의미하며, n_{11}/n_{1+} 이 된다. 신뢰도 $C(A \Rightarrow B)$ 에서 전항과 후항을 바꾸어 계산되는 신뢰도 $C(B \Rightarrow A)$ 는 항목 집합 B 가 포함된 거래 비율 중 항목 집합 A 와 항목 집합 B 가 동시에 포함된 거래의 비율 ($conf_2$)을 의미하며, n_{11}/n_{+1} 이 된다. 신뢰도는 하나의 항목 발생이 다른 하나의 항목 발생에 영향을 미치는 비율을 의미한다. 그러나 위의 정의에서 보는 바와 같이 항목 X 가 발생할 비율 $P(X)$ 는 고려하지 않으므로 바람직하지 않은 연관성 규칙이 생성될 수도 있다. 마지막으로 향상도 $L(A \Rightarrow B)$ 는 항목 집합 A 를 구매한 경우 그 거래가 항목 집합 A 를

포함하는 경우와 항목 집합 B 가 임의로 구매되는 경우의 비를 의미하며, $n_{11} \cdot n / (n_{1+} \cdot n_{+1})$ 로 계산된다. 향상도는 두 항목 간에 상관관계를 표현하고 있으며, 값이 1 보다 크면 두 항목 간에 양의 관련성이 있다고 할 수 있다. 만약 향상도의 값이 1에 가까우면 두 항목은 독립이라고 할 수 있으며, 그 값이 1 보다 작으면 두 항목은 음의 관련성이 있는 것으로 판단할 수 있다. 따라서 향상도는 신뢰도의 결점을 부분적으로 보완해주는 측도로 볼 수 있다. 그러나 일반적으로 향상도의 값이 1 이상인 경우에 의미 있는 연관성 규칙으로 간주한다 (Park, 2012).

이러한 사실을 고려해볼 때, 3가지 연관성 규칙 평가 기준값을 이용하여 연관성 규칙 수를 추정하는 문제를 연구할 필요가 있다. 이를 위해 Yi 등 (2011)은 연관성 평가 기준들 중에서 최저지지도와 최저신뢰도만을 고려하여 일반적으로 많이 활용되고 있는 다음의 세 가지 비선형 회귀 모형을 적용한 바 있다. 그리고 Park (2013)은 Yi 등 (2011)이 제안한 최저지지도와 최저신뢰도에 의한 회귀 모형뿐만 아니라 최저지지도와 최저향상도에 의한 회귀 모형, 그리고 최저신뢰도와 최저향상도에 의한 회귀 모형 등 9가지의 모형을 적용한 바 있다. 본 논문에서는 Yi 등 (2011)이 고려한 비선형회귀모형에 기본적인 연관성 평가기준 모두를 적용한 다음의 3가지 모형을 제안하고자 한다. 여기서 x_1 은 최저지지도, x_2 는 최저신뢰도, 그리고 x_3 는 최저향상도이다.

$$[\text{모형 1}] y = \beta_0 + \beta_1 \frac{1}{x_1} + \beta_2 \frac{1}{x_2} + \beta_3 \frac{1}{x_3} + \beta_4 \frac{1}{x_1^2} + \beta_5 \frac{1}{x_2^2} + \beta_6 \frac{1}{x_3^2} + \epsilon$$

$$[\text{모형 2}] y = \beta_0 + \beta_1 \frac{1}{\sqrt{x_1}} + \beta_2 \frac{1}{\sqrt{x_2}} + \beta_3 \frac{1}{\sqrt{x_3}} + \beta_4 \frac{1}{x_1} + \beta_5 \frac{1}{x_2} + \beta_6 \frac{1}{x_3} + \epsilon$$

$$[\text{모형 3}] y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 \frac{1}{x_1} + \beta_5 \frac{1}{x_2} + \beta_6 \frac{1}{x_3} + \epsilon$$

이들 3가지 모형에 대해 간단한 변수변환을 하게 되면 이들 모두 선형 모형으로 만들 수 있으며, 일반적인 최소제곱법을 이용하여 회귀계수를 추정할 수 있다. Yi 등 (2011)은 회귀식을 추정하였으나 회귀모형의 가정 특히 다중공선성 문제에 대해서는 전혀 고려하지 않고 회귀식을 추정하였다. 이에 본 논문에서는 분산팽창계수를 이용하여 다중공선성 문제를 진단하는 동시에 분산분석 결과와 수정 결정계수를 이용하여 각 모형의 기여도를 비교하여 가장 바람직한 회귀 모형을 구하고자 한다. 또한 이 모형을 Park (2013)에서 고려한 9개의 모형과 다중선형회귀모형과도 비교하고자 한다.

3. 적용 예제

본 절에서는 예제를 통하여 Yi 등 (2011)과 Park (2013)이 고려한 모형들을 포함하여 2절에서 제시한 3가지 모형에 대해 비교하고자 한다. 이를 위해 먼저 지지도와 신뢰도 및 향상도에 대한 난수를 Park (2013)과 동일하게 발생시키고자 한다. 즉, 지지도는 [0.1, 0.5]의 범위 내에서 일양 난수 (uniform random number)를, 신뢰도는 [0.3, 0.9]의 범위 내에서 일양 난수를, 향상도는 [0.7, 5.0]의 범위 내에서 일양 난수를 발생시켰으며, 이를 정리하여 Table 3.1에 제시하였다.

그 다음으로는 연관성 평가 기준의 값에 따라 연관성 규칙의 수를 생성하기 위해 지지도와 신뢰도 및 향상도의 최저값에 대한 난수를 생성하였다. 모형들의 비교를 위해 Park (2013)에서와 마찬가지로 연관성 규칙의 평가기준인 최저지지도 (m_s)는 평균이 0.35이고, 표준편차가 0.1인 정규난수 (normal random number)를 발생시키고, 최저신뢰도 (m_c)는 평균이 0.6이고, 표준편차가 0.15인 정규난수를 발생시켰으며, 최저향상도 (m_l)는 평균이 0.6이고, 표준편차가 0.9인 정규난수를 발생시킨 후, 이들을 근거로 하여 Table 3.1의 각각의 기준을 만족하는 규칙의 수를 Table 3.2에 제시하였다. 예를 들어 최저지지도와 최저신뢰도, 그리고 최저향상도의 기준을 근거로 하여 Table 3.1의 자료를 이용하여 연관성

Table 3.1 Interestingness measures by simulation

$S(A \Rightarrow B)$	conf	lift	supp	conf	lift	supp	conf	lift
0.34	0.72	0.90	0.15	0.88	2.35	0.40	0.44	3.51
0.16	0.40	0.93	0.36	0.42	2.35	0.40	0.69	3.52
0.12	0.36	1.02	0.44	0.46	2.35	0.14	0.56	3.57
0.30	0.33	1.03	0.48	0.43	2.39	0.26	0.55	3.74
0.42	0.43	1.04	0.46	0.48	2.40	0.14	0.79	3.78
0.40	0.43	1.11	0.47	0.50	2.40	0.25	0.56	3.82
0.37	0.43	1.13	0.10	0.83	2.41	0.23	0.67	3.83
0.12	0.82	1.17	0.35	0.33	2.42	0.28	0.61	3.86
0.12	0.38	1.25	0.32	0.50	2.49	0.34	0.67	3.90
0.42	0.68	1.29	0.20	0.86	2.59	0.21	0.82	3.96
0.13	0.62	1.35	0.18	0.89	2.67	0.17	0.55	4.09
0.42	0.65	1.36	0.12	0.54	2.71	0.47	0.84	4.12
0.17	0.54	1.38	0.47	0.74	2.74	0.15	0.33	4.13
0.33	0.87	1.48	0.43	0.69	2.77	0.15	0.87	4.13
0.45	0.47	1.61	0.16	0.38	2.80	0.22	0.33	4.13
0.49	0.72	1.62	0.31	0.89	2.83	0.15	0.62	4.25
0.31	0.47	1.70	0.27	0.86	2.85	0.34	0.30	4.25
0.32	0.74	1.75	0.39	0.67	2.89	0.45	0.84	4.27
0.16	0.70	1.85	0.14	0.34	2.96	0.35	0.37	4.44
0.34	0.83	1.90	0.25	0.67	3.09	0.25	0.42	4.45
0.46	0.67	2.02	0.41	0.42	3.09	0.47	0.81	4.47
0.32	0.75	2.07	0.46	0.82	3.12	0.23	0.39	4.48
0.46	0.80	2.13	0.25	0.64	3.21	0.45	0.66	4.48
0.23	0.45	2.18	0.20	0.67	3.26	0.31	0.45	4.51
0.24	0.38	2.25	0.45	0.65	3.26	0.18	0.53	4.56
0.11	0.85	2.26	0.32	0.83	3.35	0.39	0.80	4.63
0.30	0.90	2.28	0.43	0.80	3.35	0.26	0.46	4.65
0.37	0.60	2.28	0.23	0.90	3.38	0.49	0.69	4.68
0.48	0.83	2.28	0.25	0.86	3.39	0.19	0.31	4.88
0.35	0.32	2.30	0.24	0.62	3.48	0.14	0.55	4.95

규칙의 수를 생성하였을 때, $m_s = 0.38$, $m_c = 0.84$, $m_l = 2.72$ 이면 연관성 규칙의 수 n 은 7개가 생성되었으며, $m_s = 0.23$, $m_c = 0.64$, $m_l = 2.31$ 이면 연관성 규칙의 수 n 은 20개가 되며, $m_s = 0.46$, $m_c = 0.74$, $m_l = 4.55$ 이면 연관성 규칙의 수는 하나도 생성되지 않았음을 의미한다.

Table 3.2 Association thresholds and number of association rule by simulation

m_s	m_c	m_l	n	m_s	m_c	m_l	n	m_s	m_c	m_l	n
0.29	0.98	1.97	5	0.24	0.50	4.31	5	0.15	0.35	3.34	6
0.43	0.71	4.24	2	0.26	0.74	2.84	10	0.57	0.57	4.11	0
0.38	0.84	2.72	7	0.41	0.59	2.48	10	0.32	0.46	2.58	3
0.28	0.61	1.74	24	0.32	0.51	3.40	8	0.45	0.70	3.80	14
0.43	0.48	3.11	4	0.46	0.74	4.55	0	0.35	0.77	3.36	5
0.42	0.63	2.48	10	0.37	0.50	3.18	9	0.36	0.53	3.08	9
0.23	0.58	2.76	23	0.31	0.58	3.73	7	0.44	0.78	4.08	5
0.38	0.71	2.84	13	0.39	0.74	2.93	12	0.36	0.50	3.40	17
0.14	0.27	2.88	30	0.42	0.65	3.83	5	0.28	0.65	1.99	5
0.28	0.61	4.44	0	0.34	0.90	4.44	0	0.42	0.58	1.74	13
0.32	0.86	3.47	5	0.45	0.70	2.19	6	0.27	0.83	3.39	0
0.37	0.95	2.69	8	0.51	0.61	2.22	0	0.30	0.54	3.27	9
0.49	0.72	3.20	0	0.44	0.85	1.09	3	0.33	0.71	3.43	4
0.44	0.52	2.74	8	0.36	0.82	5.29	0	0.35	0.85	2.89	0
0.30	0.36	4.11	8	0.27	0.83	4.53	0	0.37	0.49	3.11	10
0.49	0.50	3.57	18	0.34	0.85	3.16	0	0.37	0.54	5.08	7
0.45	0.59	3.73	26	0.25	0.64	1.79	5	0.37	0.61	4.74	0
0.28	0.47	2.94	3	0.30	0.60	1.95	20	0.41	0.52	2.38	10
0.51	0.52	3.33	0	0.34	0.80	1.15	9	0.44	0.56	3.30	4
0.27	0.69	4.32	2	0.13	0.16	1.89	31	0.42	0.59	2.89	0
0.32	0.33	2.89	7	0.39	0.46	1.72	7	0.33	0.53	3.24	10
0.47	0.55	3.39	12	0.40	0.77	2.37	5	0.37	0.78	2.92	25
0.36	0.37	1.86	17	0.40	0.71	2.28	23	0.23	0.64	2.31	20
0.19	0.38	4.02	5	0.32	0.66	2.94	11	0.20	0.76	3.12	21
0.41	0.66	2.26	9	0.43	0.89	3.01	0	0.28	0.41	3.62	9
0.03	0.66	1.99	34	0.48	0.85	3.07	8	0.24	0.63	5.55	0
0.27	0.28	3.82	11	0.22	0.70	3.98	4	0.24	0.72	2.24	13
0.47	0.48	3.71	3	0.22	0.81	1.46	13	0.45	0.61	1.84	11
0.51	0.59	3.72	0	0.43	0.44	3.99	10	0.27	0.66	3.13	9

이들 자료를 이용하여 3가지 모형의 추정된 회귀계수와 분산분석의 유의확률, 그리고 수정된 결정계수를 나타내면 Table 3.3과 같다. 여기서 DW는 Durbin-Watson의 자기상관계수를 의미한다. 예를 들어 최소 지지도가 0.27, 최소신뢰도가 0.66, 그리고 최소향상도가 3.13이라고 할 때 모형 1을 이용하여 연관성 규칙의 수를 추정하기 위한 회귀식은 다음과 같다.

$$\hat{y} = -23.466 + 12.978 \frac{1}{x_1} - 0.215 \frac{1}{x_2} - 0.253 \frac{1}{x_3} - 0.794 \frac{1}{x_1^2} + 0.031 \frac{1}{x_2^2} + 0.919 \frac{1}{x_3^2}$$

이 식을 이용하여 추정된 연관성 규칙의 수는 13.377이 됨을 알 수 있다.

Table 3.3 Estimated regression coefficient, p-value by ANOVA, and adjusted R^2

model	β_0	β_1	β_2	β_3	β_4	β_5	β_6	p-value	Adj. R^2	DW
[1]	-23.466	12.978	-0.215	-0.253	-0.794	0.031	0.919	0.000	0.966	1.755
[2]	-57.084	48.892	0.541	-5.153	-6.070	-0.211	4.391	0.000	0.961	1.812
[3]	10.256	-37.328	-1.879	0.335	3.529	-0.469	2.447	0.000	0.955	1.885

이들 3가지 모형을 비교해 볼 때, 모든 모형이 분산분석의 유의확률값이 0.001 이하로 매우 유의하게 나타났으며, 자기상관계수도 모두 2에 가까우므로 자기 상관이 존재하지 않는다고 볼 수 있다. 또한 수정결정계수의 크기는 [모형 1]이 0.966으로 가장 크게 나타나고 있어서 이 모형에 포함된 독립변수들에 의한 회귀모형의 기여율이 가장 높다는 것을 알 수 있으나 이 모형은 분산팽창계수가 10보다 크게 나타나고 있어서 다중공선성의 문제가 있으므로 회귀모형으로는 바람직하지 않다. [모형 2]도 분산팽창계수가 10보다 크게 나타나서 다중공선성의 문제가 있다. 반면에 [모형 3]은 분산팽창계수가 10보다 작게 나타나고 있어서 다중공선성의 문제가 존재하지 않는다. 또한 이들 3가지 모형 모두 오차항의 자기상관이 존재하지 않는다고 할 수 있다. 이 결과를 종합해 볼 때, [모형 3]이 가장 바람직한 모형이라고 할 수 있다.

이 모형들을 Park (2013)의 연구에서 고려한 최저지지도, 최저신뢰도, 그리고 최저향상도의 쌍으로 구성된 회귀 모형과 비교하기 위해 연구결과를 인용하면 다음과 같다. 먼저 [모형 1]에 관계되는 3가지 형태의 비선형 회귀 모형은 다음과 같다.

$$[\text{모형 1-1}] y = \beta_0 + \beta_1 \frac{1}{x_1} + \beta_2 \frac{1}{x_2} + \beta_3 \frac{1}{x_1^2} + \beta_4 \frac{1}{x_2^2} + \epsilon$$

$$[\text{모형 1-2}] y = \beta_0 + \beta_1 \frac{1}{x_1} + \beta_2 \frac{1}{x_3} + \beta_3 \frac{1}{x_1^2} + \beta_4 \frac{1}{x_3^2} + \epsilon$$

$$[\text{모형 1-3}] y = \beta_0 + \beta_1 \frac{1}{x_2} + \beta_2 \frac{1}{x_3} + \beta_3 \frac{1}{x_2^2} + \beta_4 \frac{1}{x_3^2} + \epsilon$$

이에 대한 회귀 계수 추정 결과는 다음과 같다.

Table 3.4 Estimated regression coefficient, p-value by ANOVA, and adjusted R^2

model	β_0	β_1	β_2	β_3	β_4	p-value	Adj. R^2	DW
[1-1]	-61.990	12.467	33.377	-0.319	-3.618	0.000	0.927	2.092
[1-2]	-57.624	9.816	172.659	-0.231	-118.664	0.000	0.912	1.944
[1-3]	-85.674	32.423	241.146	-3.333	-176.663	0.000	0.872	1.947

[모형 1]에 관계되는 3가지 형태의 비선형 회귀 모형 모두 분산분석의 유의확률값이 0.001 이하로 매우 유의하게 나타났으며, 자기상관계수의 값도 모두 2에 가까우므로 자기 상관이 존재하지 않는다고 볼

수 있다. 이들 중에서는 [모형 1-1]의 수정결정계수가 0.927로 가장 크게 나타나고 있으나 이 값은 [모형 1]의 수정결정계수의 크기보다는 작다. 또한 이들 네 가지 모형 모두 분산팽창계수가 10보다 크게 나타나고 있어서 다중공선성의 문제가 있으므로 회귀모형으로는 바람직하지 않다.

[모형 2]에 관계되는 Park (2013)에서 나타난 3가지 형태의 비선형 회귀 모형은 다음과 같다.

$$[\text{모형 2-1}] \quad y = \beta_0 + \beta_1 \frac{1}{\sqrt{x_1}} + \beta_2 \frac{1}{\sqrt{x_2}} + \beta_3 \frac{1}{x_1} + \beta_4 \frac{1}{x_2} + \epsilon$$

$$[\text{모형 2-2}] \quad y = \beta_0 + \beta_1 \frac{1}{\sqrt{x_1}} + \beta_2 \frac{1}{\sqrt{x_3}} + \beta_3 \frac{1}{x_1} + \beta_4 \frac{1}{x_3} + \epsilon$$

$$[\text{모형 2-3}] \quad y = \beta_0 + \beta_1 \frac{1}{\sqrt{x_2}} + \beta_2 \frac{1}{\sqrt{x_3}} + \beta_3 \frac{1}{x_2} + \beta_4 \frac{1}{x_3} + \epsilon$$

Table 3.5 Estimated regression coefficient, p-value by ANOVA, and adjusted R^2

model	β_0	β_1	β_2	β_3	β_4	p-value	Adj. R^2	DW
[2-1]	-205.798	73.199	134.866	-8.893	-29.122	0.000	0.937	2.020
[2-2]	-175.299	53.969	291.433	-5.741	-161.284	0.000	0.906	1.998
[2-3]	-266.620	140.681	412.911	-31.219	-240.027	0.000	0.880	2.115

이 모형들을 본 논문에서 고려하는 [모형 2]와 비교해볼 때, 수정결정계수의 크기는 [모형 2]가 가장 크다는 것을 알 수 있으나 여기서도 이들 네 가지 모형 모두 분산팽창계수가 10보다 크게 나타나고 있어서 다중공선성의 문제가 있다. [모형 2]에 관계되는 3가지 형태의 비선형 회귀 모형 모두 분산분석의 유의확률값이 0.001 이하로 매우 유의하게 나타났으며, 자기상관계수의 값도 모두 2에 가까우므로 자기 상관이 존재하지 않는다고 볼 수 있다. 또한 이들 중에서는 [모형 2-1]의 수정결정계수가 0.937로 가장 크게 나타나고 있으나 [모형 2]의 수정결정계수의 크기보다는 작다. 여기서도 네 가지 모형 모두 분산 팽창계수가 10보다 크게 나타나고 있어서 다중공선성의 문제가 있으므로 회귀모형으로는 바람직하지 않다.

마지막으로 Park (2013)에서 제시된 [모형 3]에 관계되는 3가지 형태의 비선형 회귀 모형은 다음과 같다.

$$[\text{모형 3-1}] \quad y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 \frac{1}{x_1} + \beta_4 \frac{1}{x_2} + \epsilon$$

$$[\text{모형 3-2}] \quad y = \beta_0 + \beta_1 x_1 + \beta_2 x_3 + \beta_3 \frac{1}{x_1} + \beta_4 \frac{1}{x_3} + \epsilon$$

$$[\text{모형 3-3}] \quad y = \beta_0 + \beta_1 x_2 + \beta_2 x_3 + \beta_3 \frac{1}{x_2} + \beta_4 \frac{1}{x_3} + \epsilon$$

Table 3.6 Estimated regression coefficient, p-value by ANOVA, and adjusted R^2

model	β_0	β_1	β_2	β_3	β_4	p-value	Adj. R^2	DW
[3-1]	86.511	-115.307	-53.888	-0.278	4.359	0.000	0.926	1.901
[3-2]	67.909	-86.310	-8.577	0.690	5.076	0.000	0.871	2.062
[3-3]	94.400	-65.240	-11.157	0.006	2.436	0.000	0.877	2.275

이 표에서 보는 바와 같이 [모형 3]에 관계되는 3가지 형태의 비선형 회귀 모형 모두 분산분석의 유의확률값이 0.001 이하로 매우 유의하게 나타났으며, 자기상관계수도 모두 2에 가까우므로 자기 상관이

존재하지 않는다고 볼 수 있다. 이들 중에서는 [모형 3-1]의 수정결정계수가 0.926로 가장 크게 나타나고 있으나 이 값은 [모형 3]의 수정결정계수의 크기보다는 작다. 또한 이들은 모두 분산팽창계수가 10보다 작게 나타나고 있어서 다중공선성의 문제가 존재하지 않는다.

따라서 Park (2013)의 연구와 본 연구를 종합적으로 검토해볼 때, [모형 3]이 가장 바람직한 회귀모형으로 간주할 수 있다. 뿐만 아니라 [모형 3]을 Park (2013)의 연구결과인 지지도, 신뢰도, 향상도의 연관성 평가 기준의 3가지 쌍에 대한 다중선형회귀모형과도 비교해보았다.

Table 3.7 Estimated simple regression equation, p-value, and adjusted R^2

model	estimated regression equation	p-value	Adj. R^2	DW
10	$\hat{y} = 102.779 - 114.612x_1 - 69.612x_2$	0.000	0.918	1.956
11	$\hat{y} = 79.850 - 100.450x_1 - 9.486x_3$	0.000	0.859	2.182
12	$\hat{y} = 96.250 - 65.152x_2 - 11.485x_3$	0.000	0.880	2.264

이들 모형은 모두 다중공선성 문제가 없다. 그러나 이들 중에서 수정 결정계수의 최대값은 0.918로 나타나고 있어서 [모형 3]보다 더 좋은 모형을 발견할 수 없었다.

마지막으로 지지도, 신뢰도, 향상도의 연관성 평가 기준 모두에 대한 다중선형회귀모형과도 비교해보았다. 이에 대한 추정식은 $\hat{y} = 36.840 - 75.580x_1 - 3.739x_2 - 0.035x_3$ 이며, 이는 다중공선성이 존재하지 않으며, DW=2.082, 수정된 결정계수의 값은 0.920이다. 따라서 이 모형도 [모형 3]보다 더 좋다고 할 수 없다.

4. 결론

빅 데이터는 방대한 데이터의 분석을 통해 미래를 예측하고 전망하는 역할을 할 것이며, 데이터마이닝 기법에 의해 과거보다 훨씬 정확하고 과학적인 방법으로 미래를 예측할 수 있을 것이다. 이들 중에서 연관성 규칙은 기본적인 흥미도 측도인 지지도, 신뢰도, 향상도 등을 바탕으로 규칙 생성 여부를 판단하게 된다. 이들에 대한 평가기준값들을 너무 작게 하면 매우 많은 연관성 규칙이 생성되며, 너무 크게 하면 생성되는 연관성 규칙의 수가 상당히 작아지게 된다. 따라서 실제로 의미 있는 평가기준을 정하기 위해서는 반복적으로 조정 과정을 거쳐야 한다. 본 논문에서는 이러한 문제를 해결하기 위해 지지도와 신뢰도 및 향상도의 평가기준값 모두에 대해 일반적으로 많이 활용되고 있는 비선형 회귀모형들을 적용하여 연관성 규칙의 수를 추정하였다. 또한 분산팽창계수를 이용하여 다중공선성을 진단하는 동시에 수정 결정계수 및 분산분석 결과를 이용하여 각 모형의 기여도를 비교하여 가장 바람직한 회귀 모형을 구하였다. 그 결과, [모형 3]은 분산팽창계수가 10보다 작게 나타나고 있어서 다중공선성의 문제가 존재하지 않는 동시에 오차항의 자기상관이 존재하지 않는다고 할 수 있으며, 수정 결정계수의 크기를 다른 모형과 비교해도 별 차이가 없으므로 [모형 3]이 가장 바람직한 모형이라고 할 수 있다. 또한 Park (2013)이 고려한 9가지 모형을 본 논문에서 고려하는 3가지 유사한 모형으로 분류하여 비교해 본 결과, 여기서도 [모형 3] 보다 더 나은 모형을 발견할 수 없었다. 마지막으로 연관성 평가 기준 모두를 고려한 다중선형회귀모형과도 비교해보았으나 [모형 3]이 가장 바람직한 모형으로 나타났다. 따라서 기본적인 연관성 평가 기준 전부를 고려한 회귀모형이 Yi 등 (2011)과 Park (2013)이 제안한 모형 보다 바람직한 모형이라고 할 수 있다. 또한 연관성 규칙의 수를 추정하기 위해 지지도와 신뢰도 및 향상도의 평가 기준값 모두에 대해 모형을 적용할 경우, 각각의 기준과 각 기준의 역수를 고려한 비선형 회귀모형이 가장 바람직하다고 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. *Proceeding of the 20th VLDB Conference*, 487-499.
- Bayardo, R. J. (1998). Efficiently mining long patterns from databases. *Proceedings 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. (2011a). Study on the multi intervening relation in association rules. *Journal of the Korean Data Analysis Society*, **13**, 297-306.
- Cho, K. H. and Park, H. C. (2011b). Discovery of insignificant association rule s using external variable. *Journal of the Korean Data Analysis Society*, **13**, 1343-1352.
- Han, J. and Fu, Y. (1995). Discovery of multiple-level association rules from large databases. *Proceeding of the 21st VLDB Conference*, 420-431.
- 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. and Kamber, M. (2006). *Data mining ; Concepts and techniques*, Morgam Kaufmann Publishers, San Francisco.
- 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. (2011a). Proposition of negatively pure association rule threshold. *Journal of the Korean Data & Information Science Society*, **22**, 179-188.
- Park, H. C. (2011b). The proposition of attributably pure confidence in association rule mining. *Journal of the Korean Data & Information Science Society*, **22**, 235-243.
- Park, H. C. (2011c). The application of some similarity measures to association rule thresholds. *Journal of the Korean Data Analysis Society*, **13**, 1331-1342.
- Park, H. C. (2012). Exploration of symmetric similarity measures by conditional probabilities as association rule thresholds. *Journal of the Korean Data Analysis Society*, **14**, 707-716.
- Park, H. C. (2013). A study on comparison of non-linear regression model for decision of association rule numbers. *Journal of the Korean Data Analysis Society*, **15**, To be published.
- 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.
- Srinkant, R., Vu, Q. and Agrawal, R. (1997). Mining association rules with item constraints. *Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining*, 67-73.
- Toivonen, H. (1996). Sampling large database for association rules. *Proceedings of the 22nd VLDB Conference*, 134-145.
- Yi, W, Lu, M. and Liu, Z. (2011). Regression analysis in the number of association rules. *International Journal of Automation and Computing*, **8**, 78-82.

Non-linear regression model considering all association thresholds for decision of association rule numbers

Hee Chang Park¹

¹Department of Statistics, Changwon National University

Received 4 February 2013, revised 20 February 2013, accepted 1 March 2013

Abstract

Among data mining techniques, the association rule is the most recently developed technique, and it finds the relevance between two items in a large database. And it is directly applied in the field because it clearly quantifies the relationship between two or more items. When we determine whether an association rule is meaningful, we utilize interestingness measures such as support, confidence, and lift. Interestingness measures are meaningful in that it shows the causes for pruning uninteresting rules statistically or logically. But the criteria of these measures are chosen by experiences, and the number of useful rules is hard to estimate. If too many rules are generated, we cannot effectively extract the useful rules. In this paper, we designed a variety of non-linear regression equations considering all association thresholds between the number of rules and three interestingness measures. And then we diagnosed multi-collinearity and auto-correlation problems, and used analysis of variance results and adjusted coefficients of determination for the best model through numerical experiments.

Keywords: Association threshold, confidence, interestingness measure, lift, non-linear regression equations, support.

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