Journal of the Korean Data & Information Science Society
2016, **27**(3), 629–638

http://dx.doi.org/10.7465/jkdi.2016.27.3.629
한국데이터정보과학회지

연관성 규칙 수의 추정을 위한 일반적인 비선형 회귀모형에서의 표준화 향상도 활용 방안

박희창[1]

[1]창원대학교 통계학과

접수 2016년 4월 21일, 수정 2016년 5월 11일, 게재확정 2016년 5월 21일

요 약

최근에 많이 활용되고 있는 데이터 분석을 위한 연관성 규칙 마이닝은 대용량 데이터베이스에 많이 활용되고 있는 서 두 항목간의 관계를 측도화 함으로써 두 개 이상의 항목간의 관련성을 표시하여 주는 기법이다. 연관성 규칙의 여부를 판단하기 위한 연관성 평가 기준에는 지지도, 신뢰도, 그리고 향상도 등이 있으며, 이들 세 가지 기준을 이용하여 연관성 규칙 생성 여부를 판단하게 된다. 이에 대한 기존의 연구 결과는 결정함수를 이용하는 방법과 회귀모형을 이용하는 방법으로 분류할 수 있다. 회귀모형을 이용하여 수행한 연구에는 지지도와 신뢰도에 의한 모형, 세 가지 평가 기준의 쌍에 의한 모형, 표준화 향상도를 포함한 세 가지 평가 기준의 쌍에 의한 모형, 그리고 세 가지 평가 기준 전부를 고려한 모형 등이 있다. 본 논문에서는 기존의 연구를 확장하는 의미에서 표준화 향상도를 포함한 세 가지 평가 기준 전부를 고려한 비선형 회귀모형을 이용하여 연관성 규칙의 수를 추정하는 방안에 대해 강구하고자 한다. 또한 분산분석에서의 F 통계량과 수정 결정계수를 이용하여 각 모형의 유의한 정도를 비교하는 동시에 분산팽창계수에 의한 공선성 문제를 진단함으로써 가장 유용한 회귀 모형을 탐색하고자 한다.

주요용어: 신뢰도, 일반적 비선형 회귀모형, 지지도, 표준화 향상도, 흥미도 측도.

1. 서론

데이터의 급속한 증가로 인한 데이터 혁명은 이제껏 데이터와 정보가 부족하게 살아왔던 우리에게 넘쳐나는 데이터에 대한 새로운 데이터의 기술을 요구하고 있으며, 이러한 시대 변화에 대응하여 빅데이터에 대한 전반적인 이해와 이를 통한 지식의 창출 및 활용에 대한 문제의식을 갖는 것은 매우 중요하다고 할 수 있다 (Han과 Jin, 2014). 빅 데이터dp 내재된 정보를 성공적으로 얻기 위해서는 데이터 마이닝 기법의 적용이 필수적인데, 이는 방대한 양의 데이터에서 쉽게 드러나지 않는 유용한 정보를 찾아내는 과정을 의미하는 것으로, 특정 분야의 의사결정을 위한 의미 있는 정보를 확보할 수 있다 (Park, 2014b).

데이터 마이닝 분야에서 많이 고려되고 있는 연관성 규칙은 대용량 데이터베이스에서 두 항목간의 관계를 흥미도 측도 (interestingness measures)를 통하여 여러 항목들 간의 관련성을 나타내기 때문에 백화점, 인터넷 비즈니스, 통신업, 그리고 제조업계 등의 현업에서 직접 적용이 가능하다 (Park, 2013b). 이러한 연관성 규칙이 Agrawal 등 (1993)에 의해 소개된 이래로 여러 학자들이 이에 대한 연구를 진행한 바 있다 (Silberschatz 등, 1996; Lim 등 (2010); Tan 등, 2002; Geng 등, 2006; Jin 등, 2011; Park,

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

2011a; Park, 2011b; Cho와 Park, 2013; Park, 2013a; Park, 2013b; Lee와 Bae, 2014; Park, 2014a; Park, 2014b; Park, 2015).

특정 항목에 대한 연관성 규칙을 평가하기 위해 활용하는 기본적인 연관성 평가 기준에는 지지도 (support), 신뢰도 (confidence), 그리고 향상도 (lift) 등이 있다. 이들 3가지 기준에 따라 생성되는 규칙의 수가 결정되며, 특히 발생빈도가 매우 작은 항목인 경우에는 이들 모두를 충족하는 경우가 거의 없으므로 이들 측도 중 어느 하나만이라도 기준 이상이 되는 규칙에 대해 순위를 부여한 연관성 순위 결정 함수가 필요하다 (Park, 2010a). 이를 위해 Wu 등 (2004)은 이들 세 가지 평가 기준을 이용하여 연관성 규칙 생성 여부를 판단하기 위한 함수를 제안하였으며, Park (2010b)는 이들 기준 전부가 만족되지 않는 경우의 연관성 규칙을 순서화하기 위한 연관성 순위 결정 함수를 개발하는 연구를 수행하였다. 또한 Park (2010b)는 특정 연관성 평가 기준의 영향을 배제하기 위해 향상도 영향의 축소에 의한 연관성 순위 결정 함수를 제안한 바 있으며, Park (2010c)와 Park (2010d)은 각각 조건부 확률증분비와 표준화 향상도를 이용한 연관순위결정함수를 제안한 바 있다.

한편, 연관성 규칙 수의 추정 문제를 해결하기 위해 회귀분석기법을 적용한 연구가 수행되었는데, 먼저 Yi 등 (2011)은 지지도와 신뢰도에 대해 일반적으로 많이 활용되고 있는 비선형 회귀모형을 적용한 바 있다. 그러나 그들은 다중회귀모형에서 반드시 고려하여야 할 다중공선성 문제를 진단하지 않았는데 Park (2013a)는 이러한 문제를 해결하기 위해 기준값들의 값을 짝을 지워 연관성 규칙의 수를 추정하는 문제를 다루었다. Park (2014b)는 향상도는 값의 변화 구간의 범위가 상당히 넓으므로 이를 그대로 회귀모형에 적용하기에는 무리가 따른다는 점에 착안하여 향상도 대신에 표준화 향상도를 적용하였으며, Park (2013b)는 3가지 기준 전부를 적용한 비선형 회귀모형을 고려한 바 있다. 본 논문에서는 Park (2013b)와 Park (2014b)의 연구를 확장하는 의미에서 지지도, 신뢰도, 그리고 표준화 향상도를 동시에 고려한 일반적인 비선형 회귀모형을 이용하여 연관성 규칙의 수를 추정하는 방안에 대해 논의하고자 한다. 특히 모의실험을 통해 얻어진 회귀분석 결과를 이용하여 기존의 모형들에 비해 제안한 모형의 유용성을 살펴보고자 한다.

2. 표준화 향상도를 포함한 일반적인 비선형 회귀 모형의 제안

Park (2013b)는 연관성 규칙의 수를 추정하기 위해 3가지 기본적인 연관성 평가기준 모두를 반영한 비선형 회귀모형을 고려한 바 있는데 이들을 식으로 나타내면 다음과 같다.

$$supp(X \Rightarrow Y) = P(X \cap Y),\ conf(X \Rightarrow Y) = P(Y|X),\ lift(X \Rightarrow Y) = \frac{P(Y|X)}{P(Y)}$$

지지도 $supp(X \Rightarrow Y)$는 두 항목 X와 Y의 동시발생비율을 의미하며, 신뢰도 $conf(X \Rightarrow Y)$는 항목 X의 발생이 항목 Y의 발생에 영향을 미치는 비율을 의미한다. 그리고 향상도 $lift(X \Rightarrow Y)$는 두 항목 X와 Y간에 상관관계를 표현하며, 값이 1 보다 크면 이들 항목 간에는 양으로 관련성이 있다고 할 수 있다.

회귀모형을 이용하여 연관성 규칙 수를 추정하기 위해 Yi 등 (2011)은 향상도를 제외하고 지지도와 신뢰도만을 이용하였으며, 다중공선성에 관한 논의는 하지 않았다. 이를 보완하기 위해 Park (2013a)는 이들 연관규칙 평가 기준 값들을 쌍으로 고려하여 연관성 규칙의 수를 추정하는 문제를 다룬 바 있으며, Park (2013b)는 Table 2.1과 같이 평가 기준 모두를 동시에 고려한 비선형 회귀 모형을 고려하였다. 여기서 종속변수 y는 연관성 규칙의 수를 의미하며, ϵ은 오차항을 의미한다.

Table 2.1 Types of non-linear regression model by elementary thresholds

Model	Regression equation
1-1	$y = \beta_0 + \beta_1 \frac{1}{supp} + \beta_2 \frac{1}{conf} + \beta_3 \frac{1}{lift} + \beta_4 \frac{1}{supp^2} + \beta_5 \frac{1}{conf^2} + \beta_6 \frac{1}{lift^2} + \epsilon$
2-1	$y = \beta_0 + \beta_1 \frac{1}{\sqrt{supp}} + \beta_2 \frac{1}{\sqrt{conf}} + \beta_3 \frac{1}{\sqrt{lift}} + \beta_4 \frac{1}{supp} + \beta_5 \frac{1}{conf} + \beta_6 \frac{1}{lift} + \epsilon$
3-1	$y = \beta_0 + \beta_1 supp + \beta_2 conf + \beta_3 lift + \beta_4 \frac{1}{supp} + \beta_5 \frac{1}{conf} + \beta_6 \frac{1}{lift} + \epsilon$

Park (2014b)는 향상도는 다른 평가 기준에 비해 변화의 폭이 상당히 크므로 이를 그대로 회귀모형에 적용하기에는 무리가 따른다는 점에 착안하여 향상도 대신에 표준화 향상도를 적용하였으며, Park (2013b)는 3가지 기본적인 연관성 평가기준 모두를 반영한 비선형 회귀모형을 고려한 바 있다.

그러나 지지도와 신뢰도는 범위가 [0, 1]인 반면에 이 표에서 고려하는 향상도는 [0, ∞]이므로 이를 회귀모형구축에 그대로 적용하게 되면 큰 값의 영향을 받게 되어 바람직한 결과를 얻지 못할 수도 있으므로 Park (2014b)는 향상도 대신 다음과 같이 최대 및 최소값을 이용한 표준화 향상도를 지지도 및 신뢰도와 쌍으로 회귀모형에 적용한 바 있다.

$$Sm_l = \frac{lift - \min_l}{\max_l - \min_l}$$

이 식에서 $\min_l$과 $\max_l$은 각각 향상도의 최소값과 최대값을 의미한다. Park (2014b)가 고려한 모형은 Table 2.2와 같다.

Table 2.2 Non-linear models by elementary thresholds and standardized lift

Regression equation
$y = \beta_0 + \beta_1 \frac{1}{supp} + \beta_2 \frac{1}{Sm_l} + \beta_3 \frac{1}{supp^2} + \beta_4 \frac{1}{Sm_l^2} + \epsilon$
$y = \beta_0 + \beta_1 \frac{1}{conf} + \beta_2 \frac{1}{Sm_l} + \beta_3 \frac{1}{conf^2} + \beta_4 \frac{1}{Sm_l^2} + \epsilon$
$y = \beta_0 + \beta_1 \frac{1}{\sqrt{supp}} + \beta_2 \frac{1}{\sqrt{Sm_l}} + \beta_3 \frac{1}{supp} + \beta_4 \frac{1}{Sm_l} + \epsilon$
$y = \beta_0 + \beta_1 \frac{1}{\sqrt{conf}} + \beta_2 \frac{1}{\sqrt{Sm_l}} + \beta_3 \frac{1}{conf} + \beta_4 \frac{1}{Sm_l} + \epsilon$
$y = \beta_0 + \beta_1 supp + \beta_2 Sm_l + \beta_3 \frac{1}{supp} + \beta_4 \frac{1}{Sm_l} + \epsilon$
$y = \beta_0 + \beta_1 conf + \beta_2 Sm_l + \beta_3 \frac{1}{conf} + \beta_4 \frac{1}{Sm_l} + \epsilon$

본 논문에서는 Park (2013b)와 Park (2014b)의 연구를 확장하는 의미에서 연관성 규칙의 수를 지지도, 신뢰도, 그리고 표준화 향상도를 독립변수로 동시에 고려한 비선형 회귀모형을 이용하여 추정하고자 하며, 이들 모형에 대해 회귀계수의 유의성과 함께 다중 공선성, 모형의 적합도, 그리고 설명력 등과 관련된 논의를 하고자 한다. 기존 연구에서와 마찬가지로 본 논문에서도 Table 2.3과 같이 선형 가능한 비선형 회귀 모형을 고려하고자 한다.

Table 2.3 Types of non-linear regression model considering standardized lift

Model	Regression equation
1-2	$y = \beta_0 + \beta_1 \frac{1}{supp} + \beta_2 \frac{1}{conf} + \beta_3 \frac{1}{Sm_l} + \beta_4 \frac{1}{supp^2} + \beta_5 \frac{1}{conf^2} + \beta_6 \frac{1}{Sm_l^2} + \epsilon$
2-2	$y = \beta_0 + \beta_1 \frac{1}{\sqrt{supp}} + \beta_2 \frac{1}{\sqrt{conf}} + \beta_3 \frac{1}{\sqrt{Sm_l}} + \beta_4 \frac{1}{supp} + \beta_5 \frac{1}{conf} + \beta_6 \frac{1}{Sm_l} + \epsilon$
3-2	$y = \beta_0 + \beta_1 supp + \beta_2 conf + \beta_3 Sm_l + \beta_4 \frac{1}{supp} + \beta_5 \frac{1}{conf} + \beta_6 \frac{1}{Sm_l} + \epsilon$

3. 적용 예제

본 절에서는 2 종류의 예제 (case 1, case2)를 통하여 Park (2013b)가 고려한 Table 2.1의 모형과 본 논문에서 고려하는 Table 2.3의 모형에 대해 비교하고자 한다. 이를 위해 먼저 Park (2013b)와 유사하게 통계패키지인 SPSS를 이용하여 지지도는 [0.2, 0.5], 신뢰도는 [0.4, 0.9], 향상도는 [2.0, 5.0]의 범위 내에서 각각 균일 난수 (uniform random number)를 40개씩 생성하였으며, 이들 중에서 각 항목의 발생확률보다 큰 지지도는 제거한 후 최종 36개에 대한 기본적인 통계량을 구하여 Table 3.1에 제시하였다.

Table 3.1 Elementary statistics for association thresholds by case 1

Thresholds	N	Min	Max	Mean	S.D.
$supp$	36	.211	.486	.31625	.084632
$conf$	36	.424	.896	.67638	.144857
$lift$	36	2.011	4.906	3.1492	.947347

또한 연관성 평가 기준이 취하는 값의 크기에 따라 연관성 규칙의 수를 생성하기 위해 Park (2013b)와 같은 방법으로 지지도와 신뢰도 및 향상도의 최소값인 m_s, m_c, m_l에 대한 난수를 생성하였다. m_s는 평균이 0.35, 표준편차가 0.1, m_c는 평균이 0.6, 표준편차가 0.1, 그리고 m_l는 평균이 3.5이고 표준편차가 0.2인 정규난수 (normal random number) 각각 생성한 후, m_l을 표준화한 값인 Sm_l을 근거로 하여 규칙들을 생성하였다. 이에 대한 기술통계량의 값은 Table 3.2와 같다.

Table 3.2 Elementary statistics for association thresholds by case 1

Thresholds	N	Min	Max	Mean	S.D.
m_s	36	.088	.502	.34897	.095329
m_c	36	.355	.768	.60950	.079336
m_l	36	3.332	3.973	3.55661	.165002
Sm_l	36	.001	1.000	.35040	.257414

이 표에서 보는 바와 같이 최소 연관성 평가 기준을 그대로 사용하게 되면 향상도의 범위가 1을 초과하게 되는 반면에 표준화한 향상도는 0과 1 사이의 값을 취하게 되므로 세 측도 모두 범위가 0과 1 사이의 값을 갖는다.

다음으로는 기본적인 연관성 평가 기준에 의한 각 모형의 추정된 회귀계수와 회귀모형의 적합도 (F value)에 대한 유의확률, 수정된 결정계수 (Adjusted), 그리고 분산팽창계수 (VIF)를 이용하여 각 모형 간의 비교를 통하여 본 논문에서 제안하는 모형의 유용성을 고찰하고자 한다. 이를 위해 먼저 Model 1-1과 1-2의 결과를 Table 3.3에 제시하였다. 이 표에서 *는 $p < 0.05$, **는 $p < 0.01$, 그리고 ***는 $p < 0.001$을 의미한다. 이 표로부터 알 수 있는 바와 같이 두 모형의 F값이 각각 57.474와 38.361로 나타나서 두 모형 모두 적합도가 유의한 것으로 나타났으며, 수정 결정계수의 값도 각각 0.894와 0.865로 계산되어서 상당히 큰 값으로 얻어졌다. 그러나 VIF의 값을 살펴보면 두 모형 모두 공선성의 문제가 있는 것으로 나타났다.

Table 3.3 Estimated regression coefficients of model 1 by case 1

coefficient	Model			
	1-1		1-2	
	value	VIF	value	VIF
β_0	-209.675*		-18.164**	
β_1	5.601***	15.685	5.722***	15.654
β_2	8.306	58.897	8.192	57.301
β_3	1340.661*	2003.342	0.012	5921.933
β_4	-.341***	15.758	-.349***	15.710
β_5	-1.754	57.981	-1.929	56.595
β_6	-2345.195	2001.012	-6.400e-6	5920.330
F	57.474***		38.361***	
Adjusted R^2	0.894		0.865	

Model 2-1과 2-2의 추정된 회귀계수들을 계산하면 Table 3.4와 같이 얻어진다. 먼저 Model 2-1의 결과를 살펴보면 F값이 76.182로 상당히 유의하게 나타났고, 수정된 결정계수도 0.864로 크게 나타나고 있어서 모형이 상당히 의미가 있는 것으로 나타났다. 그러나 두 모형의 VIF의 값을 살펴보면 Model 2-1에 비해 Model 2-2가 많이 줄어들기는 하였으나 여전히 공선성의 문제가 나타난다고 할 수 있다.

Table 3.4 Estimated regression coefficients of model 2 by case 1

coefficient	Model			
	2-1		2-2	
	value	VIF	value	VIF
β_0	-740.519*		-72.205**	
β_1	32.806***	45.838	33.704***	45.534
β_2	46.368	232.464	50.522	228.557
β_3	2516.850	7881.272	.114	96.685
β_4	-5.716***	45.872	-5.903***	45.528
β_5	-15.502	231.010	-17.619	226.635
β_6	-2357.589	7877.147	-.002	96.204
F	52.749***		42.477***	
Adjusted R^2	0.899		0.877	

Model 3-1과 3-2의 추정된 회귀계수들을 계산한 결과를 나타내면 Table 3.5와 같다. 이 표에서 보는 바와 같이 두 모형 모두 F값이 유의한 것으로 나타났으며, 수정 결정계수의 값도 상당히 큰 값으로 얻어졌다. 또한 Model 3-1에서는 지지도의 계수를 제외하고는 모든 계수들이 유의하지 않는 것으로 나타난 반면에 Model 3-2에서는 지지도의 계수와 신뢰도의 계수가 유의한 것으로 나타났다. 공선성의 유무를 확인하기 위해 VIF의 값을 비교해보면 Model 3-1에서는 대부분의 계수들이 큰 값으로 나타나고 있어서 공선성의 문제가 존재하는 반면에 Model 3-2에서는 상대적으로 적은 것으로 나타나고 있어서 Model 3-1에 비해 공선성 문제가 심각하지 않다고 할 수 있다.

Table 3.5 Estimated regression coefficients of model 3 by case 1

coefficient	Model			
	3-1		3-2	
	value	VIF	value	VIF
β_0	337.339		39.270***	
β_1	-35.473***	3.926	-35.783***	3.920
β_2	-18.729	16.270	-21.162*	15.864
β_3	-42.763	486.297	-1.265	1.407
β_4	-.281	3.893	-.309	3.871
β_5	-4.212	16.022	-5.015	15.514
β_6	-530.427	485.381	.000	1.078
F	44.428***		40.531***	
Adjusted R^2	0.882		0.871	

이 예제를 통하여 본 논문에서 고려한 3가지 모형의 F값과 수정된 결정계수를 비교해보면 모든 모형이 의미가 있는 것으로 나타났으나 VIF의 값에 의해 공선성의 문제가 존재하는 것으로 나타났다. 이들 중에는 Model 3-2의 VIF 값이 16 이하로 나타났으므로 3가지 모형 중에서는 Model 3-2가 좀 더 바람직한 것으로 나타났다.

다음으로는 또 다른 모의실험의 결과를 이용하여 모형을 비교하기 위해 3가지 평가 기준인 지지도, 신뢰도, 향상도를 각각 [0.1, 0.3], [0.5, 0.9], [1.5, 2.2]의 범위에서 균일 난수를 100개씩 생성하였으며, 이들 중에서 각 항목의 발생확률보다 큰 지지도는 제거한 후 최종 72개에 대한 기본적인 통계량을 구하여 Table 3.6에 제시하였다.

Table 3.6 Elementary statistics for association thresholds by case 2

Thresholds	N	Min	Max	Mean	S.D.
$supp$	72	.101	.295	.19779	.055931
$conf$	72	.504	.899	.71293	.135310
$lift$	72	1.512	2.185	1.87104	.208057

또한 평가 기준을 근거로 하여 연관성 규칙의 수를 얻기 위해 m_s, m_c, m_l을 각각 평균 0.2와 표준편차 0.05, 평균 0.5와 표준편차 0.1, 그리고 평균 2.0과 표준편차 0.1인 정규난수를 생성하였다. 이에 대한 기술통계량의 값은 Table 3.7과 같다.

Table 3.7 Elementary statistics for association thresholds by case 2

Thresholds	N	Min	Max	Mean	S.D.
m_s	72	.100	.279	.17568	.036750
m_c	72	.348	.762	.54676	.090708
m_l	72	1.837	2.207	1.99418	.097688
Sm_l	72	.00	1.000	.4248	.264020

이 데이터를 이용하여 얻어진 Model 1-1과 1-2의 결과는 Table 3.8과 같다. 이 표로부터 알 수 있는 바와 같이 Model 1-1에서는 향상도와 관련된 회귀계수의 값이 다른 평가 기준에 대한 회귀계수들에 비해 매우 큰 값으로 나타났으나 전혀 유의하지 않는 것으로 나타난 반면에 Model 1-2에서는 지지도의 역제곱값에 대한 회귀계수를 제외한 모든 회귀계수의 값이 유의한 것으로 나타났다. 각 모형의 F 값에서 나타난 바와 같이 두 모형 모두 적합도가 유의한 것으로 나타났으며, 수정 결정계수의 값도 각각 0.866과 0.402로 계산되어서 상당히 큰 값으로 얻어졌다. 그러나 VIF의 값을 살펴보면 Model 1-1에 비해 Model 1-2가 적게 나타나기는 하였으나 공선성의 문제가 존재하는 것으로 나타났다.

Table 3.8 Estimated regression coefficients of model 1 by case 2

coefficient	Model			
	1-1		1-2	
	value	VIF	value	VIF
β_0	-455.890**		-107.404***	
β_1	11.425***	54.339	10.932*	54.372
β_2	66.031***	75.662	75.052***	75.904
β_3	1154.081	1525.167	.034*	4.961
β_4	-.618***	54.089	-.608	53.784
β_5	-14.660***	76.134	-17.350**	76.278
β_6	-894.529	1523.999	3.472e-5*	4.903
F	77.599***		8.946***	
Adjusted R^2	0.866		0.402	

Model 2-1과 2-2의 추정된 회귀계수들을 계산하면 Table 3.9와 같이 얻어진다. 먼저 Model 2-1의 결과를 살펴보면 F값이 76.182로 상당히 유의하게 나타났고, 수정된 결정계수도 0.864로 크게 나타나고 있어서 모형이 상당히 의미가 있는 것으로 나타났다. 반면에 Model 1-1에서와 마찬가지로 향상도와 관련된 계수들은 매우 큰 값으로 계산되었으나 전혀 유의하지 않는 것으로 나타났다. Model 2-2의 결과를 살펴보면 Model 2-1과 같이 F값이 상당히 유의하게 얻어졌고, 수정된 결정계수도 크게 나타나고 있어서 모형이 의미가 있는 것으로 나타났다. 그리고 지지도의 역제곱근값에 대한 회귀계수를 제외하고는 모든 회귀계수들의 값이 유의한 것으로 나타났다. 또한 VIF의 값을 살펴보면 Model 2-1에서는 모든 값이 매우 큰 값으로 나타난 반면에 Model 2-2에서는 모두 10보다 작거나 비슷한 값으로 나타났으므로 공선성이 존재하지 않는 것으로 해석할 수 있다.

Table 3.9 Estimated regression coefficients of model 2 by case 2

coefficient	Model			
	2-1		2-2	
	value	VIF	value	VIF
β_0	-1391.428*		-116.995***	
β_1	83.565***	200.146	24.454***	4.488
β_2	307.167***	304.092	51.426***	3.219
β_3	2614.968	6050.759	5.974***	10.052
β_4	-13.287**	199.680	-1.230	4.349
β_5	-102.380***	305.051	-2.594***	3.315
β_6	-1590.151	6048.243	-.241***	10.094
F	76.182***		27.674***	
Adjusted R^2	0.864		0.696	

Model 3-1과 3-2의 추정된 회귀계수들을 계산한 결과를 나타내면 Table 3.10과 같다. 이 표에서 보는 바와 같이 두 모형 모두 F값이 유의한 것으로 나타났으며, 수정 결정계수의 값도 상당히 큰 값으로 얻어졌다. 또한 Model 3-1에서는 지지도의 계수와 신뢰도와 관련 계수를 제외하고는 모든 계수들이 유의하지 않는 것으로 나타난 반면에 Model 3-2에서는 역의 지지도를 제외한 모든 회귀계수의 값이 유의한 것으로 나타났다. 공선성의 유무를 확인하기 위해 VIF의 값을 비교해보면 Model 3-1에서는 대부분의 계수들이 큰 값으로 나타나고 있어서 공선성의 문제가 존재하는 반면에 Model 3-2에서는 모두 3 이하로 나타나고 있어서 공선성이 없는 것으로 나타났다. Model 3-1과 Model 3-2에서는 β_1, β_2, β_3가 다른 모형과는 달리 음의 값으로 나타나는 이유는 이들 모형에서는 평가 기준의 역수 또는 역 제곱근 값이 아니라 평가 기준 그대로를 독립변수로 사용하였기 때문인 것으로 생각된다

Table 3.10 Estimated regression coefficients of model 3 by case 2

coefficient	Model			
	3-1		3-2	
	value	VIF	value	VIF
β_0	482.480		76.314***	
β_1	-154.345***	12.658	-123.114***	2.546
β_2	-92.032***	20.352	-50.584***	2.227
β_3	-120.753	377.221	-23.937***	1.038
β_4	-.812	12.549	.055	2.427
β_5	-17.649**	20.602	-1.426***	2.314
β_6	-224.221	376.555	.007*	1.034
F	72.842***		91.673***	
Adjusted R^2	0.859		0.885	

본 논문에서 고려한 3가지 모형 Model 1-2, Model 2-2, 그리고 Model 3-2에 대해 F 통계량 값과 수정 결정계수의 값을 비교해보면 Model 3-2가 가장 크게 나타났으며, VIF 값도 3 이하로 공선성이 존재하지 않는 것으로 나타났다. 따라서 3가지 모형 중에서 Model 3-2가 가장 바람직한 것으로 나타났다.

4. 결론

데이터 마이닝 기술 중의 하나인 연관성 규칙은 지지도, 신뢰도, 향상도 등의 기본적인 흥미도 측도를 기준으로 규칙의 생성 여부를 판단하게 된다. 이때 평가기준을 크게 하면 원하는 규칙의 수가 나오지 않게 되고 작게 하면 필요 이상의 연관성 규칙이 생성된다. 따라서 규칙의 수를 적절하게 하기 위해서는 평가 기준값에 대해 반복적으로 조정 과정을 거쳐야 한다 (Park, 2014a). 이는 상당히 번거로운 작업이므로 보다 단순하게 규칙의 수를 결정하기 위해 지지도와 신뢰도, 그리고 향상도의 기준값 전부에 대해

비선형 회귀모형들을 적용하여 연관성 규칙의 수를 추정할 필요가 있다. 그러나 지지도와 신뢰도에 비해 향상도는 취할 수 있는 범위가 크기 때문에 부정확한 결과를 얻을 수도 있다. 본 논문에서는 이 문제를 해결하기 위해 평가기준인 향상도 대신 표준화 향상도를 비선형 회귀모형에 적용하여 기존의 연구 결과와 비교하였다. 그 결과, 기존의 모형과 본 논문에서 제시한 모형 모두가 모형의 적합도 측면에서 유의하게 나타났으며, 기존의 모형에서는 분산팽창계수의 값이 상당히 큰 값으로 계산되어서 공선성의 문제가 심각한 반면에 본 논문에서 고려한 모형에서는 분산팽창계수의 값이 현저히 줄어들었다. 또한 본 논문에서 고려한 모형에서 유의한 회귀 계수가 더 많이 나타났다. 본 논문에서 고려한 모형 중에서는 평가 기준과 평가 기준의 역수를 고려한 모형이 가장 유용한 것으로 나타났다.

References

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*, Association for Computing Machinery, New York, USA.

Cho, K. H. and Park, H. C. (2013). A study of Gyungnam's social indicator survey using data mining. *Journal of the Korean Data Analysis Society*, **15**, 2489-2497.

Geng, L. and Hamilton, H. J. (2006). Interestingness measures for data mining: A survey. *ACM Computing Surveys*, **38**, 1-32.

Han, G. and Jin, S. (2014). Introduction to big data and the case study of its application. *Journal of the Korean Data Analysis Society*, **16**, 2447-2455.

Jin, D. S., Kang, C., Kim, K. K. and Choi, S. B. (2011). CRM on travel agency using association rules. *Journal of the Korean Data Analysis Society*, **13**, 2945-2952.

Lee, C. H. and Bae, J. H. (2014). A new importance measure of association rules using information theory. *Journal of the Korea Information Processing Society Transactions on Software and Data Engineering*, **3**, 37-42.

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**, 1061-069.

Park, H. C. (2010a). Development of associative rank decision function using basic association rule thresholds. *Journal of the Korean Data Analysis Society*, **12**, 961-972.

Park, H. C. (2010b). Association rule ranking function by decreased lift influence. *Journal of the Korean Data & Information Science Society*, **21**, 397-405.

Park, H. C. (2010c). Association rule ranking function using conditional probability increment ratio. *Journal of the Korean Data & Information Science Society*, **21**, 709-717.

Park, H. C. (2010d). Association rule ranking function using standardized lift. *Journal of the Korean Data Analysis Society*, **12**, 2661-2670.

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. (2013a). A study on comparison of non-linear regression model for decision of association rule numbers. *Journal of the Korean Data Analysis Society*, **15**, 125-132.

Park, H. C. (2013b). Non-linear regression model considering all association thresholds for decision of association rule numbers. *Journal of the Korean Data & Information Science Society*, **24**, 267-275.

Park, H. C. (2014a). Comparison of confidence measures useful for classification model building. *Journal of the Korean Data & Information Science Society*, **25**, 1-7.

Park, H. C. (2014b). Development of regression models by standardized lift for association rule number estimation. *Journal of the Korean Data Analysis Society*, **16**, 2447-2455.

Park, H. C. (2015). A study on the ordering of PIM family similarity measures without marginal probability. *Journal of the Korean Data & Information Science Society*, **26**, 367-376.

Silberschatz, A. and Tuzhilin, A. (1996). What makes patterns interesting in knowledge discovery systems. *IEEE Transactions on Knowledge Data Engineering*, **8**, 970-974.

Tan, P. N., Kumar, V. and Srivastava, J. (2002). Selecting the right interestingness measure for association patterns. *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Association for Computing Machinery, New York, USA.

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.

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.

Journal of the Korean Data & Information Science Society
2016, **27**(3), 629–638

http://dx.doi.org/10.7465/jkdi.2016.27.3.629
한국데이터정보과학회지

Generally non-linear regression model containing standardized lift for association number estimation

Hee Chang Park[1]

[1]Department of Statistics, Changwon National University

Received 21 April 2016, revised 11 May 2016, accepted 21 May 2016

Abstract

Among data mining techniques, the association rule is one of the most used in the real fields because it clearly displays the relationship between two or more items in large databases by quantifying the relationship between the items. There are three primary quality measures for association rule; support, confidence, and lift. We evaluate association rules using these measures. The approach taken in the previous literatures as to estimation of association rule number has been one of a determination function method or a regression modeling approach. In this paper, we proposed a few of non-linear regression equations useful in estimating the number of rules and also evaluated the estimated association rules using the quality measures. Furthermore we assessed their usefulness as compared to conventional regression models using the values of regression coefficients, F statistics, adjusted coefficients of determination and variation inflation factor.

Keywords: Confidence, generally non-linear regression equation, interestingness measure, standardized lift, support.

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