Journal of the Korean Data & Information Science Society
2015, **26**(4), 857–864

http://dx.doi.org/10.7465/jkdi.2015.26.4.857
한국데이터정보과학회지

부분적 주변 비율에 의한 확률적 흥미도 측도 기반 유사성 측도의 상한 및 하한의 설정

박희창[1]

[1]창원대학교 통계학과

접수 2015년 6월 11일, 수정 2015년 6월 26일, 게재확정 2015년 7월 1일

요 약

데이터 마이닝은 다양한 형태의 방대한 데이터 집합으로부터 보이지 않는 지식이나 새로운 법칙을 발견한 후, 이를 바탕으로 의사결정 등을 위한 정보로 활용하고자 하는 것이다. 데이터 마이닝 기법 중의 하나인 군집 분석은 거리 또는 유사성 측도를 이용하여 집단을 분류하고, 구분된 각 집단의 특성을 파악하기 위한 기법이다. 본 논문에서는 주변 확률이 일부 포함된 확률적 흥미도 측도 기반의 유사성 측도들인 Peirce Ⅰ, Peirce Ⅱ, Cole Ⅰ, Cole Ⅱ, 그리고 이들을 응용한 Park Ⅰ 및 Park Ⅱ에 대한 대소 관계를 수식의 증명뿐만 아니라 예제 데이터에 의해서도 규명하였다. 그 결과, Cole Ⅰ과 Cole Ⅱ의 측도를 동시에 고려한 Loevinger 측도가 기존의 측도들 중에서는 상한이 되나 Park Ⅰ 및 Park Ⅱ를 함께 고려했을 경우에는 동시발생비율, 동시 비발생비율, 그리고 두 가지 형태의 불일치비율의 크기에 따라 변한다는 사실을 확인하였다.

주요용어: 군집 분석, 부분적 주변 확률, 유사성 측도, 확률적 흥미도 측도.

1. 서론

오늘날 방대한 양의 정형 또는 비정형 데이터를 수집하고 저장할 수 있는 능력이 급격히 향상됨에 따라 데이터 마이닝 방법론 (data mining methodology)에 대한 중요도가 더욱 증대되고 있다. 데이터 마이닝은 다양한 형태의 데이터 집합으로부터 숨겨진 지식이나 새로운 법칙과 관계를 발견하고 이를 바탕으로 의사결정 등을 위한 정보로 활용하고자 하는 것이다 (Park, 2014b). 위키 백과사전에 의하면 데이터 마이닝 기법 중의 하나인 군집분석 (cluster analysis)은 주어진 데이터들의 특성을 고려하여 데이터 집단을 정의하고 유사도가 높은 데이터들을 같은 그룹으로 분류하여 데이터의 분포나 패턴을 찾아내는 기법이다. 이러한 군집분석 기법은 데이터들 간의 거리 또는 유사성에 의하여 군집을 형성하고 형성된 군집의 특성을 파악하여 군집들 사이의 관계를 분석하게 되며, 이러한 유사성 측도들은 군집화 문제를 해결하는 데 매우 중요한 역할을 하고 있다 (Park, 2015). 군집분석에 관한 최근 연구로는 Warrens (2008), Choi 등 (2010), Lee와 Kim (2011), Yeo (2011), Park (2012, 2014a, 2014b, 2015), Lim과 Lim (2012), Park과 Kim (2013), Ryu와 Park (2013) 등이 있다. 특히 Park (2012)에서는 부분적 주변 비율 (partially marginal proportion; PMP)을 고려한 확률적 흥미도 측도 (probabilistic interestingness measure; PIM)를 기반으로 하는 유사성 측도들에 대해 연관성 평가 기준으로서의 적용 가능 여부를 탐색한 바 있다. 본 논문에서는 수식 및 예제를 통해 이들과 이들을 응용한 측도들에 대해 대소 관계를 규

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

명함으로써 이들의 상한 및 하한을 설정하는 문제를 고려하고자 한다. PIM은 유사성의 정도와 방향을 동시에 나타내주는 측도이므로 많은 연구에서 활용된 바 있다. 특히 Orchard (1975)는 PIM을 이용한 Boolean Analyzer 알고리즘을 제시한 바 있으며, Imberman 등 (2002)은 머리 외상 데이터에 PIM을 적용한 바 있다. 논문의 2절에서는 이들 측도들을 소개하는 동시에 이들에 대한 대소 관계를 규명하며, 3절에서는 예제에 의해 얻어진 결과를 이용하여 측도들 간의 경계에 대해 논의한 후, 4절에서 결론을 내리고자 한다.

2. 부분적 주변 비율을 고려한 유사성 측도의 상하한의 설정

본 절에서는 PIM 기반 유사성 측도 중에서 주변비율이 부분적으로 나타나는 유사성 측도들을 수식으로 나타내기 위해 Table 2.1의 분할표를 이용하고자 한다.

Table 2.1 2×2 contingency table by proportions

		Y		Total
		1	0	
X	1	a	b	p_1
	0	c	d	q_1
Total		p_2	q_2	1

이 표에서 각 항은 $a = P(X \cap Y)$, $b = P(X \cap Y^c)$, $c = P(X^c \cap Y)$, $d = P(X^c \cap Y^c)$을 의미하며, 주변 확률은 각각 $p_1 = a + b$, $q_1 = c + d$, $p_2 = a + c$, 그리고 $q_2 = b + d$가 된다. 이로부터 PIM은 다음과 같이 정의할 수 있다. 여기서 n 전체 발생 빈도수를 의미한다.

$$PIM = n^2(ad - bc)$$

PIM은 $ad - bc$의 값에 따라 연관성의 방향과 강도를 알 수 있는 동시에 연관 정도의 순위까지도 알 수 있는 장점이 있다 (Park, 2012). 그러나 주변 비율의 크기를 고려하지 않는 경우에는 적절하지 못한 결론에 다다를 수 있으므로 본 절에서는 Warrens (2008)에 의해 정리되고 Choi 등 (2010) 및 Park (2012)에 의해 활용된 바 있는 PMP를 고려한 PIM 계열 유사성 측도를 고려하여 이들의 상한 및 하한을 설정하고자 한다. 이들 측도에는 Peirce (1884)의 S_{P1} 및 S_{P2}, Cole (1949)의 S_{C1} 및 S_{C2}, 그리고 Loevinger (1947, 1948)에 의해 제안되고 Mokken (1971), Sijtsma와 Molenaar (2002)에 의해 논의된 바 있는 S_{Lo} 등이 있다. 이들 다섯 가지 측도들을 살펴보면 분자는 동일하고 분모는 두 개의 주변비율을 고려하여 각각 다르게 나타내고 있다. 이에 추가하여 이들 측도에는 없는 p_1p_2 및 q_1q_2를 분모로 하는 두 가지 측도를 편의상 S_{Pa1}과 S_{Pa2}로 표기하고 이들 모두를 Table 2.1의 기호를 이용하여 수식으로 나타내면 다음과 같다.

Table 2.2 PIM based similarity measures with PMP

similarity measure	formula
Peirce Ⅰ	$S_{P1} = \frac{ad-bc}{p_1q_1}$
Peirce Ⅱ	$S_{P2} = \frac{ad-bc}{p_2q_2}$
Cole Ⅰ	$S_{C1} = \frac{ad-bc}{p_1q_2}$
Cole Ⅱ	$S_{C2} = \frac{ad-bc}{p_2q_1}$
Loevinger	$S_{Lo} = \frac{ad-bc}{\min(p_1q_2,\ p_2q_1)}$
Park Ⅰ	$S_{Pa1} = \frac{ad-bc}{p_1p_2}$
Park Ⅱ	$S_{Pa2} = \frac{ad-bc}{q_1q_2}$

이러한 부분적으로 주변 비율을 포함하는 측도들을 살펴보면 분자는 PIM의 형태로 나타나는 반면에 분모는 측도들마다 각기 다른 형태의 주변 비율로 나타나고 있다. 측도 S_{P1}의 분모는 $P(X=1)$과 $P(X=0)$, 측도 S_{P2}의 분모는 $P(Y=1)$과 $P(Y=0)$, 측도 S_{C1}의 분모는 $P(X=1)$과 $P(Y=0)$, 측도 S_{C2}의 분모는 $P(X=0)$과 $P(Y=1)$, S_{Pa1}의 분모는 $P(X=1)$과 $P(Y=1)$, 측도 S_{Pa2}의 분모는 $P(X=0)$과 $P(Y=0)$으로 구성되어 있다. 여기서 측도 S_{Lo}는 조건에 따라 S_{C1} 또는 S_{C2}와 동일한 형태로 나타나므로 S_{C1}과 S_{C2}를 통해서 특성을 파악할 수 있다. 이 표를 통하여 알 수 있는 바와 같이 이들 측도들 간의 대소 비교는 어느 정도는 가능하나 구체적으로 나타내기 곤란하므로 이들 유사성 측도들에 대한 상한 및 하한을 설정하는 문제를 고려하고자 한다. 이를 위해 먼저 S_{P1}과 기존의 측도의 차이를 계산하면 다음과 같다.

$$S_{P1}-S_{P2}=\frac{ad-bc}{p_1q_1p_2q_2}[p_2q_2-p_1q_1]=\frac{ad-bc}{p_1q_1p_2q_2}[(a-d)(b-c)]$$

$$S_{P1}-S_{C1}=\frac{ad-bc}{p_1q_1p_2q_2}(q_2-q_1)=\frac{ad-bc}{p_1q_1p_2q_2}(b-c)$$

$$S_{P1}-S_{C2}=\frac{ad-bc}{p_1q_1p_2}(p_2-p_1)=\frac{ad-bc}{p_1q_1p_2}(c-b)$$

따라서 $a\geq d$이고, $b\geq c$이면 $|S_{P1}|\geq|S_{P2}|$ 성립한다는 것을 알 수 있다. 또한 $b\geq c$ 이면 $|S_{P1}|\geq|S_{C1}|$이고, $c\geq b$이면 $|S_{P1}|\geq|S_{C2}|$가 성립한다.

S_{P2}과 기존의 다른 측도들과의 크기를 비교하기 위해 두 측도들 간의 차이를 계산하면 다음과 같다.

$$S_{P2}-S_{C1}=\frac{ad-bc}{p_1p_2q_2}(p_1-p_2)=\frac{ad-bc}{p_1p_2q_2}(b-c)$$

$$S_{P2}-S_{C2}=\frac{ad-bc}{p_2q_1q_2}(q_1-q_2)=\frac{ad-bc}{p_2q_1p_2}(c-b)$$

따라서 $b\geq c$이면 $|S_{P2}|\geq|S_{C1}|$이 성립하고, $c\geq b$이면 $|S_{P2}|\geq|S_{C2}|$ 가 성립한다. 또한 S_{C1}과 S_{C2}의 차이를 계산하면 다음과 같이 정리되므로 $c\geq b$이면 $|S_{C1}|\geq|S_{C2}|$가 성립한다.

$$S_{C1}-S_{C2}=\frac{ad-bc}{p_1q_1p_2q_2}[p_2q_1-p_1q_2]=\frac{ad-bc}{p_1q_1p_2q_2}(c-b)$$

다음으로는 본 논문에서 추가적으로 제안하는 두 개의 측도 S_{Pa1}과 S_{Pa2}에 대해 기존의 측도들과 대소 관계를 알아보고자 한다. 먼저 S_{Pa1}과 다른 측도들 간의 차이를 구하면 다음과 같다.

$$S_{Pa1}-S_{Pa2}=\frac{ad-bc}{p_1p_2q_1q_2}(q_1q_2-p_1p_2)=\frac{ad-bc}{p_1p_2q_1q_2}(d-a)$$

$$S_{Pa1}-S_{P1}=\frac{ad-bc}{p_1p_2q_1}(q_1-q_2)=\frac{ad-bc}{p_2q_1p_2}(d-a)$$

$$S_{Pa1}-S_{P2}=\frac{ad-bc}{p_1p_2q_2}(q_2-p_1)=\frac{ad-bc}{p_1p_2q_2}(d-a)$$

$$S_{Pa1}-S_{C1}=\frac{ad-bc}{p_1p_2q_2}(q_2-p_2)=\frac{ad-bc}{p_1p_2q_2}(b+d-a-c)$$

$$S_{Pa1} - S_{C2} = \frac{ad-bc}{p_1p_2q_1}(q_1 - p_1) = \frac{ad-bc}{p_1p_2q_2}(c+d-a-b)$$

따라서 $a \geq d$이면 $|S_{Pa1}| \geq |S_{Pa2}|$, $|S_{P1}| \geq |S_{Pa1}|$, 그리고 $|S_{P2}| \geq |S_{Pa1}|$의 관계가 성립한다. 또한 $b+d \geq a+c$이면 $|S_{Pa1}| \geq |S_{C1}|$ 이 되고, $c+d \geq a+b$이면 $|S_{Pa1}| \geq |S_{C2}|$가 된다. 다음으로는 S_{Pa2}와 다른 측도들에 대해 대소 관계를 알아보면 다음과 같은 식이 얻어진다.

$$S_{Pa2} - S_{P1} = \frac{ad-bc}{p_1q_1q_2}(p_1 - q_2) = \frac{ad-bc}{p_1q_1q_2}(a-d)$$

$$S_{Pa2} - S_{P2} = \frac{ad-bc}{p_2q_1q_2}(p_2 - q_1) = \frac{ad-bc}{p_2q_1q_2}(a-d)$$

$$S_{Pa2} - S_{C1} = \frac{ad-bc}{p_1q_1q_2}(p_1 - q_1) = \frac{ad-bc}{p_1q_1q_2}(a+b-c-d)$$

$$S_{Pa2} - S_{C2} = \frac{ad-bc}{p_2q_1q_2}(p_2 - q_2) = \frac{ad-bc}{p_2q_1q_2}(a+c-b-d)$$

따라서 $a \geq d$이면 $|S_{Pa2}| \geq |S_{P1}|$과 $|S_{Pa2}| \geq |S_{P2}|$의 관계가 성립하며, $a+b \geq c+d$이면 $|S_{Pa2}| \geq |S_{C1}|$이 되고, $a+c \geq b+d$이면 $|S_{Pa2}| \geq |S_{C2}|$ 된다. 이러한 결과들을 종합하면 Table 2.3과 같다.

Table 2.3 Bounds of PIM based measures with PMP by conditions

condition	upper and lower bounds
$a \geq d$ & $b \geq c$ & $a+b \geq c+d$	$\|S_{C1}\| \leq \|S_{Pa1}\| \leq \|S_{P2}\| \leq \|S_{P1}\| \leq \|S_{C2}\| \leq \|S_{Pa2}\|$
$a \geq d$ & $b \geq c$ & $a+b \leq c+d$	$\|S_{C1}\| \leq \|S_{Pa1}\| \leq \|S_{P2}\| \leq \|S_{P1}\| \leq \|S_{Pa2}\| \leq \|S_{C2}\|$
$a \geq d$ & $b \leq c$ & $a+c \geq b+d$	$\|S_{C2}\| \leq \|S_{Pa1}\| \leq \|S_{P1}\| \leq \|S_{P2}\| \leq \|S_{Pa2}\| \leq \|S_{C1}\|$
$a \geq d$ & $b \leq c$ & $a+c \leq b+d$	$\|S_{Pa1}\| \leq \|S_{Pa2}\| \leq \|S_{C2}\| \leq \|S_{P1}\| \leq \|S_{P2}\| \leq \|S_{C1}\|$
$a \leq d$ & $b \geq c$ & $a+b \geq c+d$	$\|S_{C1}\| \leq \|S_{Pa2}\| \leq \|S_{P1}\| \leq \|S_{P2}\| \leq \|S_{C2}\| \leq \|S_{Pa1}\|$
$a \leq d$ & $b \geq c$ & $a+b \leq c+d$	$\|S_{C1}\| \leq \|S_{Pa2}\| \leq \|S_{P1}\| \leq \|S_{P2}\| \leq \|S_{Pa1}\| \leq \|S_{C2}\|$
$a \leq d$ & $b \leq c$ & $a+c \geq b+d$	$\|S_{C2}\| \leq \|S_{Pa2}\| \leq \|S_{P2}\| \leq \|S_{P1}\| \leq \|S_{Pa1}\| \leq \|S_{C1}\|$
$a \leq d$ & $b \leq c$ & $a+c \leq b+d$	$\|S_{Pa2}\| \leq \|S_{C2}\| \leq \|S_{P2}\| \leq \|S_{P1}\| \leq \|S_{C1}\| \leq \|S_{Pa1}\|$

이러한 사실로부터 여러 가지 조건에 따라 주변 비율을 고려한 PIM 기반 유사성 측도들의 대소 관계를 규명할 수 있는 동시에 이들 측도에 대한 상한 및 하한의 값을 구할 수 있다. 특히 측도 S_{Lo}의 절대값은 모든 측도들의 절대값의 상한으로 나타나고 있다.

3. 예제를 통한 대소 관계 비교

이 절에서는 예제를 이용하여 주변 비율을 부분적으로 고려한 PIM 기반 유사성 측도들의 변화하는 패턴을 고찰하고자 한다. 이를 위해 먼저 두 항목 X와 Y간의 동시발생비율 a, 불일치비율 b 및 c, 그리고 동시 비발생비율 d의 값의 변화에 따른 여러 가지 모의실험 자료를 이용하여 측도들의 대소 관계를 살펴보는 것이 필요하다. 먼저 일치하는 방향의 비율을 나타내는 a와 d값을 각각 0.51에서 0.70, 0.01에서 0.20까지 0.01씩 증가시키면서 PMP가 포함된 유사성 측도들의 계산 결과를 나타내면 Table 3.1과 같다. 여기서 $n = 100$, $p_1 = 0.8$, $p_2 = 0.7$, $q_1 = 0.2$, $q_2 = 0.3$이며, P_{PIM}는 PIM을 전체 발생 빈도수 n의 제곱으로 나눈 값, 즉 $ad - bc$를 의미한다.

Table 3.1 Trends of similarity measures with PMP for increasing a and d

a	b	c	d	P_{PIM}	S_{C1}	S_{C2}	S_{P1}	S_{P2}	S_{Pa1}	S_{Pa2}
0.51	0.29	0.19	0.01	-0.05	-0.2083	-0.3571	-0.3125	-0.2381	-0.0893	-0.8333
0.52	0.28	0.18	0.02	-0.04	-0.1667	-0.2857	-0.2500	-0.1905	-0.0714	-0.6667
0.53	0.27	0.17	0.03	-0.03	-0.1250	-0.2143	-0.1875	-0.1429	-0.0536	-0.5000
0.54	0.26	0.16	0.04	-0.02	-0.0833	-0.1429	-0.1250	-0.0952	-0.0357	-0.3333
0.55	0.25	0.15	0.05	-0.01	-0.0417	-0.0714	-0.0625	-0.0476	-0.0179	-0.1667
0.56	0.24	0.14	0.06	0.00	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0.57	0.23	0.13	0.07	0.01	0.0417	0.0714	0.0625	0.0476	0.0179	0.1667
0.58	0.22	0.12	0.08	0.02	0.0833	0.1429	0.1250	0.0952	0.0357	0.3333
0.59	0.21	0.11	0.09	0.03	0.1250	0.2143	0.1875	0.1429	0.0536	0.5000
0.60	0.20	0.10	0.10	0.04	0.1667	0.2857	0.2500	0.1905	0.0714	0.6667
0.61	0.19	0.09	0.11	0.05	0.2083	0.3571	0.3125	0.2381	0.0893	0.8333
0.62	0.18	0.08	0.12	0.06	0.2500	0.4286	0.3750	0.2857	0.1071	1.0000
0.63	0.17	0.07	0.13	0.07	0.2917	0.5000	0.4375	0.3333	0.1250	1.1667
0.64	0.16	0.06	0.14	0.08	0.3333	0.5714	0.5000	0.3810	0.1429	1.3333
0.65	0.15	0.05	0.15	0.09	0.3750	0.6429	0.5625	0.4286	0.1607	1.5000
0.66	0.14	0.04	0.16	0.10	0.4167	0.7143	0.6250	0.4762	0.1786	1.6667
0.67	0.13	0.03	0.17	0.11	0.4583	0.7857	0.6875	0.5238	0.1964	1.8333
0.68	0.12	0.02	0.18	0.12	0.5000	0.8571	0.7500	0.5714	0.2143	2.0000
0.69	0.11	0.01	0.19	0.13	0.5417	0.9286	0.8125	0.6190	0.2321	2.1667
0.70	0.10	0.00	0.20	0.14	0.5833	1.0000	0.8750	0.6667	0.2500	2.3333

이 표에서 보는 바와 같이 a와 d 의 값이 증가하고 b와 c의 값이 감소함에 따라 P_{PIM}의 값이 증가하므로 본 논문에서 고려하는 PMP가 포함된 모든 유사성 측도들이 증가하는 것으로 나타났다. 이 표에서는 $a \geq d$, $b \geq c$, 그리고 $a+b \geq c+d$이므로 측도 S_{C2}의 절대값이 기존의 측도들 중에서는 가장 크나 S_{Pa2} 보다는 작은 것으로 나타났으며, 그 다음으로는 S_{P1}, S_{P2}, S_{Pa1}, 그리고 S_{C1}의 순으로 나타나고 있다. 이를 좀 더 구체적으로 살펴보기 위해 a, b, c, d의 값이 0.53, 0.27, 0.17, 0.03인 경우와 0.59, 0.21, 0.11, 0.09인 경우를 비교해보면 전자는 P_{PIM}의 값이 -0.03인 음의 값으로 나타난 반면에 후자의 경우에는 0.03인 양의 값으로 나타났다. 또한 각각의 경우에 대해 S_{Pa2}는 -0.5000과 0.5000, S_{C2}는 -0.2143과 0.2143, S_{P1}는 -0.1875와 0.1875, S_{P2}는 -0.1429와 0.1429, S_{Pa1}는 -0.0536과 0.0536, 그리고 S_{C1}은 -0.1250과 0.1250의 순으로 계산되었으므로 2절에서 유도한 대소 관계식이 성립된다는 것을 알 수 있다. 또한 본 논문에서 고려하는 PMP를 포함하는 모든 유사성 측도들이 전자의 경우에는 음의 값으로 나타난 반면에 후자의 경우에는 양의 값으로 나타났으며, 모든 측도들의 절대값의 크기는 두 경우가 동일한 것으로 계산되었다.

이번에는 불일치하는 방향의 비율을 나타내는 b 및 c의 값을 각각 0.02에서 0.20, 0.51에서 0.70까지 0.02씩 증가시키면서 유사성 측도들의 변화하는 경향을 나타내면 Table 3.2와 같다. 이 표에서 $n = 100$, $p_1 = 0.3$, $p_2 = 0.8$, $q_1 = 0.7$, $q_2 = 0.2$이다.

Table 3.2 Trends of similarity measures with PMP for increasing b and c

a	b	c	d	P_{PIM}	S_{C1}	S_{C2}	S_{P1}	S_{P2}	S_{Pa1}	S_{Pa2}
0.29	0.01	0.51	0.19	0.05	0.8333	0.0893	0.2381	0.3125	0.2083	0.3571
0.28	0.02	0.52	0.18	0.04	0.6667	0.0714	0.1905	0.2500	0.1667	0.2857
0.27	0.03	0.53	0.17	0.03	0.5000	0.0536	0.1429	0.1875	0.1250	0.2143
0.26	0.04	0.54	0.16	0.02	0.3333	0.0357	0.0952	0.1250	0.0833	0.1429
0.25	0.05	0.55	0.15	0.01	0.1667	0.0179	0.0476	0.0625	0.0417	0.0714
0.24	0.06	0.56	0.14	0.00	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0.23	0.07	0.57	0.13	-0.01	-0.1667	-0.0179	-0.0476	-0.0625	-0.0417	-0.0714
0.22	0.08	0.58	0.12	-0.02	-0.3333	-0.0357	-0.0952	-0.1250	-0.0833	-0.1429
0.21	0.09	0.59	0.11	-0.03	-0.5000	-0.0536	-0.1429	-0.1875	-0.1250	-0.2143
0.20	0.10	0.60	0.10	-0.04	-0.6667	-0.0714	-0.1905	-0.2500	-0.1667	-0.2857
0.19	0.11	0.61	0.09	-0.05	-0.8333	-0.0893	-0.2381	-0.3125	-0.2083	-0.3571
0.18	0.12	0.62	0.08	-0.06	-1.0000	-0.1071	-0.2857	-0.3750	-0.2500	-0.4286
0.17	0.13	0.63	0.07	-0.07	-1.1667	-0.1250	-0.3333	-0.4375	-0.2917	-0.5000
0.16	0.14	0.64	0.06	-0.08	-1.3333	-0.1429	-0.3810	-0.5000	-0.3333	-0.5714
0.15	0.15	0.65	0.05	-0.09	-1.5000	-0.1607	-0.4286	-0.5625	-0.3750	-0.6429
0.14	0.16	0.66	0.04	-0.10	-1.6667	-0.1786	-0.4762	-0.6250	-0.4167	-0.7143
0.13	0.17	0.67	0.03	-0.11	-1.8333	-0.1964	-0.5238	-0.6875	-0.4583	-0.7857
0.12	0.18	0.68	0.02	-0.12	-2.0000	-0.2143	-0.5714	-0.7500	-0.5000	-0.8571
0.11	0.19	0.69	0.01	-0.13	-2.1667	-0.2321	-0.6190	-0.8125	-0.5417	-0.9286
0.10	0.20	0.70	0.00	-0.14	-2.3333	-0.2500	-0.6667	-0.8750	-0.5833	-1.0000

이 표에서는 b와 c의 값이 증가하고 a와 d의 값이 감소하고 있으므로 P_{PIM}의 값은 감소하며, 동시에 본 논문에서 고려하는 모든 유사성 측도들도 감소하는 경향을 나타내고 있다. 이 표에서는 $a \geq d$, $b \leq c$, 그리고 $a + c \geq b + d$의 관계가 성립하므로 측도 S_{Lo}와 S_{C1}이 같은 값을 가지는 동시에 이들의 절대값이 가장 큰 것으로 나타났으며, 그 다음으로는 $|S_{Pa2}|$, $|S_{P2}|$, $|S_{P1}|$, $|S_{Pa1}|$, 그리고 $|S_{C2}|$의 순으로 나타나고 있다. 이는 Table 2.3의 세 번째 관계식이 성립된다는 것을 알 수 있다. 이러한 결과를 구체적으로 알아보기 위해 a, b, c, d의 값이 0.25, 0.05, 0.55, 0.15인 경우와 0.23, 0.07, 0.57, 0.13인 경우를 비교해보면 전자는 P_{PIM}의 값이 0.01의 양의 값으로 나타난 반면에 후자의 경우에는 -0.01의 음의 값으로 나타났다. 또한 각각의 경우에 대해 S_{C1}은 0.1667과 -0.1667, S_{Pa2}는 0.0714와 -0.0714, S_{P2}는 0.0625와 -0.0625, S_{P1}는 0.0476과 -0.0476, S_{Pa1} 0.0417과 -0.0417, 그리고 S_{C2}는 0.0179와 -0.0179의 순으로 계산되었으므로 2절에서 유도한 대소 관계식이 성립된다는 것을 알 수 있다. 또한 본 논문에서 고려하는 PMP를 포함하는 모든 유사성 측도들이 전자의 경우에는 양의 값으로 나타난 반면에 후자의 경우에는 음의 값으로 나타났으며, 모든 측도들의 절대값의 크기는 두 경우가 동일한 것으로 계산되었다. 이들 측도들은 각 경계에 있는 측도와는 유사한 값을 가지게 되므로 각 측도의 상한 및 하한은 여러 가지 측도들을 분류하는 도구로 활용될 수 있다 (Park, 2015). 또한 Warrens (2008)이 언급한 바와 같이 실제 값의 관점에서 각 측도들의 관계를 알게 되면 데이터 분석 시 이들 측도들에 대해서는 같거나 유사한 결과를 얻을 수 있으므로 주어진 알고리즘의 안정화에 도움이 될 수 있을 것으로 판단된다.

4. 결론

다양한 형태의 데이터베이스로부터 정보를 파악하기 위한 군집 분석은 거리 또는 유사성 측도를 이용하여 집단을 분류하고, 구분된 각 집단의 특성을 파악하기 위한 데이터 마이닝 기법이다. 본 논문에서는 주변 확률이 부분적으로 포함된 확률적 흥미도 측도 기반 유사성 측도들인 Peirce Ⅰ, Peirce Ⅱ, Cole Ⅰ, Cole Ⅱ, 그리고 이들을 응용한 Park Ⅰ 및 Park Ⅱ에 대한 상한 및 하한을 계산함으로써 각각의 측도에 대한 대소 관계를 수식의 증명뿐만 아니라 예제 데이터에 의해서도 규명하였다. 그 결과, Loevinger가 제안한 측도 S_{Lo}의 절대값은 기존의 측도들 중에서는 상한이 되나 Park Ⅰ 및 Park Ⅱ를 함께 고려했을 경우에는 동시발생비율, 동시 비발생비율, 그리고 두 가지 형태의 불일치비율의 크기에 따라 변한다는 사실을 확인하였다. 또한 기존의 측도들 간의 크기를 비교해 보았을 때 두 항목간의 불일치비율 b가 c보다 큰 경우, 동시발생비율 a가 동시 비발생비율 d보다 크면 측도들의 절대값의 크기가 $|S_{C2}|$, $|S_{P1}|$, $|S_{P2}|$, $|S_{C1}|$의 순으로 나타난 반면에 a가 d보다 작으면 다른 것은 모두 동일하고 $|S_{P1}|$와 $|S_{P2}|$의 위치만 바뀌었다. 만약 b가 c보다 작은 경우, a가 d보다 크면 측도들의 절대값의 크기가 $|S_{C1}|$, $|S_{P2}|$, $|S_{P1}|$, $|S_{C2}|$의 순으로 나타난 반면에 a가 d보다 작으면 이 경우에도 다른 것은 모두 동일하고 $|S_{P1}|$와 $|S_{P2}|$ 위치만 바뀌게 된다는 사실을 확인하였다.

References

Choi, S. S., Cha, S. H. and Tappert, C. (2010). A survey of binary similarity and distance measures. *Journal on Systemics, Cybernetics and Informatics*, **8**, 43-48.

Cole, L. C. (1949). The measurement of interspecific association. *Ecology*, **30**, 411-424.

Imberman S., Domanski B. and Tho mpson H. (2002), Using dependency/association rules to find indications for computerized tomography in a head trauma dataset. *Artificial Intelligence in Medicine*, **26**, 55-68.

Lee, K. A. and Kim, J. H. (2011). Comparison of clustering with yeast microarray gene expression data. *Journal of the Korean Data & Information Science Society*, **22**, 741-753.

Lim, J. S. and Lim, D. H. (2012). Comparison of clustering methods of microarray gene expression data. *Journal of the Korean Data & Information Science Society*, **23**, 39-51.

Loevinger, J. A. (1947). A systematic approach to the construction and evaluation of test ability. *Psychological Monograph*, **61**, 1-49.

Loevinger, J. A. (1948). The technique of homogeneous tests compared with some aspects of scale analysis and factor analysis. *Psychological Bulletin*, **45**, 507-529.

Mokken, R. J. (1971). *A Theory and Procedure of Scale Analysis*, The Hague, Netherlands.

Orchard R. A. (1975). *On the determination of relationships between computer system state variables*, Bell Laboratories Technical Memorandum, Bell Laboratories, New Jersey.

Park, H. C. (2012). Exploration of PIM based similarity measures with PMP as association rule thresholds. *Journal of the Korean Data Analysis Society*, **14**, 2965-2974.

Park, H. C. (2014a). Comparison of cosine family similarity measures in the aspect of association rule. *Journal of the Korean Data Analysis Society*, **16**, 729-737.

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

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.

Park, H. J. and Kim, J. T. (2013). Classification of universities in Daegu·Gyungpook by support vector cluster analysis. *Journal of the Korean Data & Information Science Society*, **24**, 783-791.

Peirce, C. S. (1884). The numerical measure of the success of predictions. *Science*, **4**, 453-454.

Ryu, J. Y. and Park, H. C. (2013). A study on Jaccard dissimilarity measures for negative association rule generation. *Journal of the Korean Data Analysis Society*, **15**, 3111-3121.

Sijtsma, K. and Molenaar, I. W. (2002). *Introduction to Nonparametric Item Response Theory*, Thousand Oaks, Sage.

Warrens, M. J. (2008). *Similarity coefficients for binary data : Properties of coefficients, coefficient Matrices, multi-way metrics and multivariate coefficients*, Doctoral d issertation, Leiden university, Netherlands.

Yeo, I. K. (2011). Clustering analysis of Korea's meteorological data. *Journal of the Korean Data & Information Science Society*, **22**, 941-949.

Journal of the Korean Data & Information Science Society
2015, **26**(4), 857–864

http://dx.doi.org/10.7465/jkdi.2015.26.4.857
한국데이터정보과학회지

Bounds of PIM-based similarity measures with partially marginal proportion

Hee Chang Park[1]

[1]Department of Statistics, Changwon National University

Received 11 June 2015, revised 26 June 2015, accepted 1 July 2015

Abstract

By Wikipedia, data mining is the computational process of discovering patterns in huge data sets involving methods at the intersection of association rule, decision tree, clustering, artificial intelligence, machine learning. Clustering or cluster analysis is the task of grouping a set of objects in such a way that objects in the same group are more similar to each other than to those in other groups. The similarity measures being used in the clustering may be classified into various types depending on the characteristics of data. In this paper, we computed bounds for similarity measures based on the probabilistic interestingness measure with partially marginal probability such as Peirce Ⅰ, Peirce Ⅱ, Cole Ⅰ, Cole Ⅱ, Loevinger, Park Ⅰ, and Park Ⅱ measure. We confirmed the absolute value of Loevinger measure wasthe upper limit of the absolute value of any other existing measures. Ordering of other measures is determined by the size of concurrence proportion, non-simultaneous occurrence proportion, and mismatch proportion.

Keywords: Cluster analysis, partially marginal probability, probabilistic interestingness measure, similarity measure.

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