Journal of the Korean Data & Information Science Society
2015, **26**(2), 367–376

http://dx.doi.org/10.7465/jkdi.2015.26.2.367
한국데이터정보과학회지

주변 확률을 고려하지 않는 확률적 흥미도 측도 계열 유사성 측도의 서열화

박희창[1]

[1]창원대학교 통계학과

접수 2015년 2월 10일, 수정 2015년 3월 4일, 게재확정 2015년 3월 18일

요 약

데이터마이닝 기법 중의 하나인 군집분석은 다양한 특성을 지닌 관찰대상에 대해 유사성을 바탕으로 동질적인 군집으로 묶은 후, 동일 군집에 속해 있는 공통된 특성을 조사하는데 이용되는 기법이다. 본 논문에서는 주변 확률을 고려하지 않는 확률적 흥미도 측도 기반 유사성 측도인 Yule I과 II, Michael, Digby, Baulieu, 그리고 Dispersion 측도에 대해 상한 및 하한을 설정함으로써 이들의 대소 관계를 규명하였다. 그 결과, 세 가지 유형의 대소 관계가 성립한다는 사실을 수식의 증명뿐만 아니라 실제 데이터 및 모의실험 데이터에 의해서도 확인할 수 있었다. 이들 측도들은 각 경계에 있는 측도와는 더욱 더 유사한 값을 가지므로 각 측도의 상한 및 하한은 여러 가지 측도들을 분류하는 도구가 되며, 실제 값의 관점에서 각 측도들의 관계를 알게 되면 주어진 알고리즘의 안정화에 도움이 될 수 있을 것이다.

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

1. 서론

최근 IT 분야의 화두는 단연 빅 데이터 (big data)이다. 현재와 과거의 PC 또는 노트북 사양을 비교해보면 데이터가 기하급수적으로 늘어났다는 사실을 실감할 수 있다. 특히 스마트 기기의 급속한 보급과 인터넷 및 소셜 미디어 등으로 대표되는 다양한 정보 채널의 등장으로 인하여 데이터는 폭발적으로 증가하고 있다. 빅 데이터에 대해 다양한 분야에서 여러 학자들이 정의하고 있는데 Lee (2013)가 정리한 바와 같이 기존의 관리 방법이나 분석 체계로는 처리하기 어려운 다양하고도 방대한 양의 정형 또는 비정형 데이터로부터 수집, 검색, 분석을 신속하게 처리하여 경제적인 가치발굴을 수행하도록 설계된 차세대 기술 및 아키텍처라고 정의할 수 있다. 이러한 빅 데이터로부터 정보를 채굴해내는 데이터 마이닝 기법 중의 하나인 군집분석 (cluster analysis)은 각 객체의 클래스 레이블이 알려지지 않은 데이터 객체의 집합을 유사성 측도에 기초하여 군집 안의 객체끼리는 높은 유사성을 지니고, 다른 군집들의 객체와는 매우 다르도록 데이터를 그룹화 시키는 과정을 말한다 (Kim 등, 2010). 이러한 군집화는 패턴 인식, 이미지 처리, 시장조사를 포함한 많은 응용 분야에서 넓게 사용된다는 점에서 매우 흥미로운 연구 분야이다 (Park, 2014).

군집분석에서는 속성 변수의 특징에 따라 크게 수치형, 범주형, 혼합형 데이터로 나누어지는데 이 때 이용되는 유사성 측도는 이러한 데이터의 속성에 따라 여러 가지의 형태로 분류할 수 있으며, 범주형

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

변수에 대한 속성값의 유사성은 값의 순서가 고유하게 정해지지 않아서 정의하기가 어렵다 (Stanfill과 Waltz, 1986; Kim 등, 2010). 유사성 측도와 관련된 군집분석에 관한 최근 연구로는 Warrens (2008), Choi 등 (2010), Lee와 Kim (2011), Yeo (2011), Park (2012, 2014), Lim과 Lim (2012), Park과 Kim (2013), Ryu와 Park (2013) 등이 있다. 본 논문에서는 이분형 데이터에 대해 주변 확률 (marginal probability; MP)을 고려하지 않는 확률적 흥미도 측도 (probabilistic interestingness measure; PIM) 기반 유사성 측도들에 대해 대소 관계를 규명함으로써 이들의 상한 및 하한을 설정하는 문제를 고려하고자 한다. 이를 위해 Warrens (2008)에 의해 정리되고 Choi 등 (2010) 및 Park (2012)에 의해 활용된 바 있는 MP를 고려하지 않는 PIM 계열 유사성 측도인 Yule I과 II, Michael, Digby, Baulieu, 그리고 Dispersion 측도들을 동시에 비교하고자 한다. 논문의 2절에서는 이들 측도들을 소개하는 동시에 이들에 대한 대소 관계를 규명하며, 3절에서는 실제 예제와 모의실험을 통한 결과를 이용하여 상한 및 하한을 살펴본 후, 4절에서 결론을 내리고자 한다.

2. 주변 확률을 고려하지 않는 유사성 측도의 대소 관계

2.1. PIM 계열 유사성 측도

본 절에서는 PIM 기반 유사성 측도 중에서 원래의 공식에서 MP가 존재하지 않는 유사성 측도를 MP를 고려하지 않는 PIM 계열 유사성 측도라고 명명하기로 한다. 이러한 유사성 측도들을 수식으로 나타내기 위해 Park (2014)에서와 같이 Table 2.1의 분할표를 이용하고자 한다.

Table 2.1 2×2 contingency table by proportions

		B		Total
		1	0	
A	1	a	b	$a+b$
	0	c	d	$c+d$
Total		$a+c$	$b+d$	n

이 표에서 각 항은 $a = n \times P(A \cap B)$, $b = n \times P(A \cap B^c)$, $c = n \times P(A^c \cap B)$, $d = n \times P(A^c \cap B^c)$을 의미하며, 이로부터 PIM은 다음과 같이 정의할 수 있다.

$$PIM = ad - bc$$

Park (2012)에서 기술한 바와 같이 PIM은 $ad - bc$의 값에 따라 연관성의 방향과 강도를 알 수 있는 동시에 연관 정도의 순위까지도 알 수 있는 장점이 있다. Warrens (2008)에 의해 정리되고 Choi 등 (2010) 및 Park (2012)에 의해 활용된 바 있는 MP를 고려하지 않는 PIM 계열 유사성 측도에는 Yule (1900)의 S_{Yule1}, Yule (1912)의 S_{Yule2}, Michael (1920)의 S_{Mich}, Digby (1983)의 S_{Digby}, Baulieu (1989)의 S_{Bau}, 그리고 Gordon (1999)의 S_{Disp}등이 있으며, 이들을 Table 2.1의 기호를 이용하여 수식으로 나타내면 다음과 같다.

Table 2.2 PIM family similarity measures without MP

similarity measure	formula
Yule I	$S_{Yule1} = \dfrac{ad-bc}{ad+bc}$
Yule II	$S_{Yule2} = \dfrac{\sqrt{ad}-\sqrt{bc}}{\sqrt{ad}+\sqrt{bc}}$
Michael	$S_{Mich} = \dfrac{4(ad-bc)}{(a+d)^2+(b+c)^2}$
Digby	$S_{Digby} = \dfrac{(ad)^{3/4}-(bc)^{3/4}}{(ad)^{3/4}+(bc)^{3/4}}$
Baulieu	$S_{Bau} = \dfrac{4(ad-bc)}{(a+b+c+d)^2}$
Dispersion	$S_{Disp} = \dfrac{ad-bc}{(a+b+c+d)^2}$

이러한 유사성 측도들은 분류나 군집화 등과 같은 패턴 분석 문제를 해결하는 데 있어서 매우 중요한 역할을 하고 있으므로 오랜 기간에 걸쳐 많은 연구자들이 가장 의미 있는 이분형 유사성 측도들을 찾기 위해 노력을 기울이고 있다 (Choi 등, 2010). 이들 측도들 중에서 S_{Yule1}, S_{Yule2}, 그리고 S_{Digby}의 측도들은 분모와 분자 모두 승산비(odds ratio)의 비선형 변환에 의한 측도들로써 상관계수와 같이 범위가 $[-1,\ 1]$이 되도록 변형한 것이다 (Warrens, 2008). 반면에 S_{Mich}, S_{Bau}, 그리고 S_{Disp}는 각각의 분자를 승산비를 구성하고 있는 ad와 bc의 차이로 나타내고 있다. 특히 S_{Bau}와 S_{Disp}의 분모는 표본 크기의 제곱이고, S_{Mich}의 분모는 같은 방향의 도수 크기의 합의 제곱으로 나타내고 있어서 S_{Bau}보다 큰 값으로 나타나는 경향이 있다. 이들 측도들은 범주형 자료의 군집화에는 이용 가능하나 연속형 자료에 대해서는 정밀도가 떨어지므로 이용의 한계가 있다고 볼 수 있다. 연관성 규칙 관점에서 볼 때, 본 논문에서 고려하고자 하는 측도들은 기존의 연관규칙 평가 기준과는 달리 교차표의 모든 항을 고려한 PIM의 값을 이용하여 연관성의 강도를 나타내는 측도이다. 이들 수식에서 나타나는 a와 d는 A, B 두 항목이 모두 발생하거나 두 항목 모두 발생하지 않는 경우의 빈도수를 나타내므로 두 항목의 연관성의 방향이 동일하다고 할 수 있다. 반면에 b와 c는 A, B 두 항목 중에서 하나는 발생하고 다른 하나는 발생하지 않는 경우의 빈도수를 의미하므로 두 항목의 연관성의 방향이 동일하지 않다고 할 수 있다 (Park, 2012). 따라서 ad의 값이 bc의 값에 비해 크면 양의 연관성이 있다고 할 수 있으며, 그 반대이면 음의 연관성이 있다고 할 수 있다.

2.2. MP를 고려하지 않는 PIM 계열 유사성 측도 비교

이제 $a+b+c+d$를 n으로 두고 이들 유사성 측도들에 대한 상한 및 하한을 설정하고자 한다. 측도 S_{Yule1}과 다른 측도들과의 대소 관계를 규명하기 위해 먼저 S_{Yule1}과 S_{Yule2}를 비교해보면 Warrens (2008)에서 증명한 바와 같이 $ad \geq bc$이면 $S_{Yule1} \geq S_{Yule2}$이고, $ad \leq bc$이면 $S_{Yule1} \leq S_{Yule2}$이므로 $|S_{Yule1}|$은 $|S_{Yule2}|$의 상한 (upper bound)이 된다. $(a+d)^2+(b+c)^2 \geq 4(ad+bc)$으로부터 $(a-d)^2+(b-c)^2 \geq 0$이 성립하므로 $|S_{Yule1}|$은 $|S_{Mich}|$의 상한 (upper bound)이 된다. S_{Yule1}과 S_{Digby}의 크기를 비교하기 위해 이들 측도의 차이를 계산하면 다음과 같다. 따라서 $ad \geq bc$이면 $S_{Yule1} \geq S_{Digby}$이고, $ad \leq bc$이면 $S_{Yule1} \leq S_{Digby}$이므로 $|S_{Yule1}|$은 $|S_{Digby}|$의 상한이 된다.

$$S_{Yule1} - S_{Digby} = 2(abcd)^{3/4}\left[(ad)^{1/4} - (bc)^{1/4}\right]$$

S_{Yule1}과 S_{Bau}의 크기를 비교하기 위해 이들 측도의 차이를 계산하면 다음과 같다. 따라서 $ad \geq bc$이면 $S_{Yule1} \geq S_{Bau}$이고, $ad \leq bc$이면 $S_{Yule1} \leq S_{Bau}$이므로 $|S_{Bau}|$는 $|S_{Yule1}|$의 하한이 된다.

$$S_{Yule1} - S_{Bau} = \frac{(ad-bc)}{n^2(ad+bc)}[(a-d)^2 + (b-c)^2 + 2(a+d)(b+c)]$$

S_{Yule1}과 S_{Disp}의 크기를 비교하기 위해 이들의 분모를 비교해보면 S_{Disp}의 분모가 S_{Yule1}의 분모보다 더 크므로 $|S_{Disp}| \leq |S_{Yule1}|$이 성립한다.

다음으로는 S_{Yule2}와 다른 측도와의 대소관계를 비교하고자 한다. 먼저 S_{Yule2}의 분모와 분자에 $\sqrt{ad}+\sqrt{bc}$를 곱하여 $S^*_{Yule2} = \frac{ad-bc}{(\sqrt{ad}+\sqrt{bc})^2}$으로 만든 후 S_{Mich}와의 차이를 계산하면 다음과 같이 정리된다.

$$S^*_{Yule2} - S_{Mich} = \frac{4(ad-bc)}{(\sqrt{ad}+\sqrt{bc})^2[(a+d)^2+(b+c)^2]}\left[\left(\frac{a+d}{2}\right)^2 + \left(\frac{b+c}{2}\right)^2 - (\sqrt{ad}+\sqrt{bc})^2\right] \tag{2.1}$$

따라서 $ad \geq bc$이고 a와 d, 그리고 b와 c의 산술평균의 제곱합이 그들의 기하평균의 합의 제곱보다 큰 경우와 $ad \leq bc$이고 이들의 산술평균의 제곱합이 그들의 기하평균의 합의 제곱보다 작은 경우에는 $S_{Yule2} \geq S_{Mich}$이 되고, 이외의 경우에는 $S_{Yule2} \leq S_{Mich}$이 된다. S_{Yule2}와 S_{Digby}의 크기를 비교하기 위해 그 차이를 계산하면 다음과 같다.

$$S_{Digby} - S_{Yule2} = 2(abcd)^{1/2}\left[(ad)^{1/4} - (bc)^{1/4}\right] \tag{2.2}$$

따라서 $ad \geq bc$이면 $S_{Digby} \geq S_{Yule2}$이고, $ad \leq bc$이면 $S_{Digby} \leq S_{Yule2}$이므로 $|S_{Yule2}|$는 $|S_{Digby}|$의 하한이 된다.

S_{Yule2}와 S_{Bau}의 크기를 비교하기 위해 이들 측도의 차이를 계산하면 다음과 같다.

$$S_{Yule2} - S_{Bau} = (\sqrt{ad} - \sqrt{bc})\left[\left(\frac{a+d}{2} + \frac{b+c}{2}\right)^2 - (\sqrt{ad} + \sqrt{bc})^2\right] \tag{2.3}$$

이 식에서 산술평균의 합이 그들의 기하평균의 합보다 항상 크므로 $ad \geq bc$이면 $S_{Yule2} \geq S_{Bau}$이고, $ad \leq bc$이면 $S_{Yule2} \leq S_{Bau}$이므로 $|S_{Bau}|$는 $|S_{Yule2}|$의 하한이 된다.

S_{Yule2}와 S_{Disp}의 크기를 비교하기 위해 이들의 차이를 계산하면 다음과 같다.

$$S_{Yule2} - S_{Disp} = (\sqrt{ad} - \sqrt{bc})\left[4\left(\frac{a+d}{2} + \frac{b+c}{2}\right)^2 - (\sqrt{ad} + \sqrt{bc})^2\right] \tag{2.4}$$

위 식에서와 마찬가지로 산술평균의 합이 그들의 기하평균의 합보다 항상 크므로 $ad \geq bc$이면 $S_{Yule2} \geq S_{Disp}$이고, $ad \leq bc$이면 $S_{Yule2} \leq S_{Disp}$이므로 $|S_{Disp}|$는 $|S_{Yule2}|$의 하한이 된다.

다음으로는 S_{Mich}와 다른 측도와의 대소 관계를 비교하기 위해 먼저 S_{Digby}와의 차이를 계산하면 다음과 같다.

$$S_{Mich} - S_{Digby} = \frac{G_1^2 - G_2^2}{A_1^2 + A_2^2} - \frac{(ad)^{3/4} - (bc)^{3/4}}{(ad)^{3/4} + (bc)^{3/4}} \tag{2.5}$$

여기서 G_1과 G_2는 각각 a와 d 및 b와 c의 기하평균이고, A_1과 A_2는 각각 a와 d 및 b와 c의 산술평균을 의미한다., 그리고 U와 V는 각각 $(ad)^{3/4}$과 $(bc)^{3/4}$을 의미한다. 따라서 $(G_1^2 - G_2^2)(U+V) \geq (A_1^2 + A_2^2)(U-V)$이면 $S_{Mich} \geq S_{Digby}$가 된다.

또한 $n^2 \geq (a+d)^2 + (b+c)^2$으로부터 $|S_{Bau}|$는 $|S_{Mich}|$의 하한 (lower bound)이 되고, Table 2.2로부터 $|S_{Disp}| \leq |S_{Mich}|$이 된다.

다음으로 S_{Digby}와 나머지 측도와의 차이를 알아보기 위해 먼저 S_{Bau}와의 관계를 알아보면 식 (2.2)에 의해 $|S_{Yule2}| \leq |S_{Digby}|$이고, 식 (2.3)에 의해 $|S_{Bau}| \leq |S_{Yule2}|$이므로 $|S_{Bau}|$는 $|S_{Digby}|$의 하한이 된다. 이와 더불어 식 (2.4)에 의해 $|S_{Disp}| \leq |S_{Yule2}|$이므로 $|S_{Disp}| \leq |S_{Digby}|$가 된다. 마지막으로 $|S_{Disp}| \leq |S_{Bau}|$가 성립한다는 사실은 Table 2.2의 수식으로부터 알 수 있다.

식 (2.1)에 의한 S_{Yule2}와 S_{Mich}의 관계 및 식 (2.5)에 의한 S_{Mich}와 S_{Digby}의 관계를 제외하면 다음과 같은 측도들 간의 관계가 항상 성립하게 된다.

$$|S_{Disp}| \leq |S_{Mich}| \leq |S_{Yule1}|$$
$$|S_{Disp}| \leq |S_{Bau}| \leq |S_{Yule2}| \leq |S_{Digby}| \leq |S_{Yule1}|$$

이에 S_{Yule2}와 S_{Mich}, 그리고 S_{Digby}의 관계를 추가하면 다음과 같이 세 가지 유형의 대소 관계로 요약할 수 있다.

[유형 1] $|S_{Disp}| \le |S_{Bau}| \le |S_{Yule2}| \le |S_{Mich}| \le |S_{Digby}| \le |S_{Yule1}|$

[유형 2] $|S_{Disp}| \le |S_{Bau}| \le |S_{Yule2}| \le |S_{Digby}| \le |S_{Mich}| \le |S_{Yule1}|$

[유형 3] $|S_{Disp}| \le |S_{Bau}| \le |S_{Mich}| \le |S_{Yule2}| \le |S_{Digby}| \le |S_{Yule1}|$

이로부터 각 조건에 따라 주변 확률을 고려하지 않는 PIM 기반 유사성 측도들의 대소 관계를 규명할 수 있는 동시에 이들 측도에 대한 경계값을 구할 수 있다.

3. 적용 예제

이 절에서는 예제를 이용하여 주변 확률을 고려하지 않은 PIM 계열 유사성 측도들이 변화하는 양상을 고찰하고자 한다. 이를 위해 먼저 2003년 경남사회지표조사 자료를 이용하였는데, 여러 가지 조사 문항들 가운데 도민의식부문 문항들 간의 유사성을 계산하여 Table 3.1에 제시하였다. 여기서 $X1$은 환경정책 시급과제로 주민의 환경의식 개혁 및 강화, $X2$는 환경사범에 대한 벌칙 강화, $X3$은 수질개선 등 환경보호시설 확충, $X4$는 민간 환경단체의 조직 및 기능 강화, $X5$는 자녀의 환경오염 저감 행동, $X6$은 1회용품 사용 유무, $X7$은 산림의 중요성 체감, $X8$은 사회복지정책 만족도, 그리고 $X9$는 환경의식 개혁 필요성을 의미한다. 이 표에서 보는 바와 같이 모든 사례에서 S_{Yule1}이 다른 측도들의 상한이 되며, 항상 $|S_{Disp}| \le |S_{Mich} \le |S_{Yule1}|$ 및 $|S_{Disp}| \le |S_{Bau}| \le |S_{Yule2}| \le |S_{Digby}| \le |S_{Yule1}|$의 관계가 성립한다. $X2$와 $X5$의 문항에서 유사성 측도들을 비교해보면 [유형 3]의 대소 관계가 성립하는데, 이는 $ad = 866,880$, $bc = 666,207$이 되어 $ad - bc = 200610$이 되어 0보다 큰 값을 가지며, 산술평균은 각각 $A_1 = 2,789$ 및 $A_2 = 831.5$, 기하평균은 각각 $G_1 = 931.1$ 및 $G_2 = 816.2$이 되어 산술평균의 제곱합= 8469913.25, 기하평균의 합의 제곱=3053057.29이 되므로 산술평균의 제곱합이 기하평균의 합의 제곱보다 크다. 따라서 식 (2.1)로부터 $S_{Mich} \le S_{Yule2}$이 성립한다. 또한 $U = 28,409.9$, $V = 23318.8$이 되므로 $(U-V)/(U+V) = 0.0984$, $(G_1^2 - G_2^2)/(A_1^2 + A_2^2) = 0.0237$이 되어 $(G_1^2 - G_2^2)(U+V)$이 $(A_1^2 + A_2^2)(U-V)$보다 작은 값으로 나타났으므로 $S_{Mich} \le S_{Digby}$가 된다.

Table 3.1 Outputs of PIM family similarity measures without MP of survey data

no.	a	b	c	d	S_{Disp}	S_{Bau}	S_{Mich}	S_{Digby}	S_{Yule2}	S_{Yule1}
$X1$: $X5$	592	3,480	558	2,611	-0.0076	-0.0302	-0.0596	-0.0853	-0.0570	-0.1136
$X1$: $X6$	2,086	3,322	1,864	2,721	-0.0052	-0.0207	-0.0413	-0.0326	-0.0218	-0.0435
$X1$: $X7$	221	5,185	419	4,165	-0.0125	-0.0502	-0.0989	-0.3113	-0.2115	-0.4048
$X1$: $X8$	1,450	3,954	1,313	3,270	-0.0045	-0.0181	-0.0360	-0.0340	-0.0227	-0.0453
$X1$: $X9$	530	4,870	545	4,039	-0.0052	-0.0206	-0.0409	-0.0805	-0.0537	-0.1071
$X2$: $X5$	160	673	990	5,418	0.0038	0.0153	0.0237	0.0984	0.0657	0.1308
$X2$: $X6$	469	733	3,481	5,310	-0.0006	-0.0025	-0.0048	-0.0091	-0.0061	-0.0121
$X2$: $X7$	113	1,089	527	8,261	0.0036	0.0144	0.0198	0.1804	0.1210	0.2386
$X2$: $X8$	362	839	2,401	6,385	0.0030	0.0119	0.0212	0.0515	0.0344	0.0686
$X2$: $X9$	146	1,056	929	7,853	0.0017	0.0066	0.0097	0.0584	0.0390	0.0778
$X3$: $X5$	210	1,263	940	4,828	-0.0033	-0.0132	-0.0229	-0.0591	-0.0394	-0.0788
$X3$: $X6$	791	1,255	3,159	4,788	-0.0018	-0.0071	-0.0140	-0.0171	-0.0114	-0.0229
$X3$: $X7$	134	1,911	506	7,439	0.0003	0.0012	0.0019	0.0114	0.0076	0.0152
$X3$: $X8$	594	1,452	2,169	5,772	0.0028	0.0112	0.0208	0.0318	0.0212	0.0424
$X3$: $X9$	263	1,782	812	7,127	0.0043	0.0172	0.0279	0.0967	0.0646	0.1287
$X4$: $X5$	43	187	1,107	5,904	0.0009	0.0036	0.0051	0.0764	0.0510	0.1017
$X4$: $X6$	115	204	3,835	5,839	-0.0011	-0.0044	-0.0086	-0.0572	-0.0382	-0.0763
$X4$: $X7$	25	294	615	9,056	0.0005	0.0018	0.0022	0.0841	0.0562	0.1120
$X4$: $X8$	95	223	2,668	7,001	0.0007	0.0028	0.0048	0.0418	0.0279	0.0557
$X4$: $X9$	38	281	1,037	8,628	0.0004	0.0015	0.0019	0.0442	0.0295	0.0589

이와 같이 실제 예제를 통해 확인하는 데에는 한계가 있으므로 두 항목 X와 Y간의 동시발생빈도 $a = n(X = 1,\ Y = 1)$, 불일치빈도 $b = n(X = 1,\ Y = 0)$와 $c = n(X = 0,\ Y = 1)$, 그리고 동시 비발생 빈도 $d = n(X = 0,\ Y = 0)$의 값의 변화에 따른 여러 가지 모의실험 자료를 이용하여 좀 더 구체적으로 주변 확률을 고려하지 않는 PIM 기반 유사성 측도들의 대소 관계를 살펴보는 것이 필요하다. 이를 위해 먼저 양으로 같은 방향의 빈도를 나타내는 a 및 d값의 증가에 따른 유사성 측도의 계산결과를 나타내면 Table 3.2와 같다. 이 표에서 보는 바와 같이 a와 d의 값이 증가함에 따라 본 논문에서 고려하는 모든 유사성 측도들이 증가하는 것으로 나타났다. 모든 경우에 S_{Yule1}의 절대값이 가장 크게 나타나고 있어서 모든 측도들의 상한이 되며, S_{Disp}의 절대값이 가장 작은 것으로 나타나서 모든 측도들의 하한이 된다. S_{Yule2}와 S_{Digby}는 항상 $|S_{Yule2}| \le |S_{Digby}|$의 관계가 성립하며, S_{Bau}, S_{Yule2}, S_{Digby}의 대소 관계는 항상 $|S_{Bau}| \le |S_{Yule2}| \le |S_{Digby}|$이 성립한다는 사실을 확인할 수 있다. 반면에 S_{Mich}와 S_{Yule2}, 그리고 S_{Digby}와 S_{Mich}의 대소 관계는 여러 가지 조건에 따라 달라짐을 알 수 있다. 이러한 사실을 좀 더 구체적으로 확인하기 위해 먼저 ($a = 3$, $b = 47$, $c = 27$, $d = 23$)의 경우와 ($a = 8$, $b = 42$, $c = 22$, $d = 28$)의 경우를 비교해보면 전자는 $S_{Yule1} = -0.897$, $S_{Digby} = -0.798$, $S_{Mich} = -0.780$, $S_{Yule2} = -0.622$, $S_{Bau} = -0.480$, $S_{Disp} = -0.120$, 그리고 후자는 $S_{Yule1} = -0.610$, $S_{Digby} = -0.486$, $S_{Mich} = -0.519$, $S_{Yule2} = -0.340$, $S_{Bau} = -0.280$, $S_{Disp} = -0.070$으로 나타나서 두 경우 모두 $|S_{Disp}| \le |S_{Mich}| \le |S_{Yule1}|$ 및 $|S_{Disp}| \le |S_{Bau}| \le |S_{Yule2}| \le |S_{Digby}| \le |S_{Yule1}|$의 관계가 성립하고 있다는 사실을 확인할 수 있다. 반면에 전자의 경우는 [유형 1]의 형태로 나타나고 있고, 후자의 경우에는 [유형 2]의 형태로 나타나고 있다.

Table 3.2 Outputs of similarity measures for increasing a and d

a	b	c	d	S_{Yule1}	S_{Digby}	S_{Mich}	S_{Yule2}	S_{Bau}	S_{Disp}
3	47	27	23	-0.897	-0.798	-0.780	-0.622	-0.480	-0.120
4	46	26	24	-0.851	-0.738	-0.737	-0.558	-0.440	-0.110
5	45	25	25	-0.800	-0.677	-0.690	-0.500	-0.400	-0.100
6	44	24	26	-0.743	-0.615	-0.637	-0.445	-0.360	-0.090
7	43	23	27	-0.679	-0.552	-0.581	-0.392	-0.320	-0.080
8	42	22	28	-0.610	-0.486	-0.519	-0.340	-0.280	-0.070
9	41	21	29	-0.535	-0.420	-0.454	-0.290	-0.240	-0.060
10	40	20	30	-0.455	-0.352	-0.385	-0.240	-0.200	-0.050
21	29	9	41	0.535	0.420	0.454	0.290	0.240	0.060
22	28	8	42	0.610	0.486	0.519	0.340	0.280	0.070
23	27	7	43	0.679	0.552	0.581	0.392	0.320	0.080
24	26	6	44	0.743	0.615	0.637	0.445	0.360	0.090
25	25	5	45	0.800	0.677	0.690	0.500	0.400	0.100
26	24	4	46	0.851	0.738	0.737	0.558	0.440	0.110
27	23	3	47	0.897	0.798	0.780	0.622	0.480	0.120
54	26	16	4	-0.316	-0.241	-0.162	-0.156	-0.080	-0.020
55	25	15	5	-0.154	-0.116	-0.077	-0.076	-0.040	-0.010
56	24	14	6	0.000	0.000	0.000	0.000	0.000	0.000
57	23	13	7	0.143	0.108	0.072	0.074	0.040	0.010
58	22	12	8	0.275	0.208	0.140	0.145	0.080	0.020
59	21	11	9	0.394	0.302	0.205	0.212	0.120	0.030
60	20	10	10	0.500	0.390	0.268	0.276	0.160	0.040
61	19	9	11	0.594	0.472	0.329	0.335	0.200	0.050
62	18	8	12	0.676	0.548	0.389	0.390	0.240	0.060
63	17	7	13	0.746	0.619	0.448	0.441	0.280	0.070

이번에는 ($a = 57$, $b = 23$, $c = 13$, $d = 7$)의 경우와 ($a = 63$, $b = 17$, $c = 7$, $d = 13$)의 경우를 비교해볼 때 앞의 예에서와 마찬가지로 두 경우 모두 $|S_{Disp}| \le |S_{Mich}| \le |S_{Yule1}|$ 및 $|S_{Disp}| \le |S_{Bau}| \le$

$|S_{Yule2}| \leq |S_{Digby}| \leq |S_{Yule1}|$의 관계가 성립하고 있으며, 전자의 경우는 [유형 3]의 형태로 나타나고 있고, 후자의 경우에는 [유형 1]의 형태로 나타난다는 사실을 확인할 수 있다.

이번에는 음으로 같은 방향을 나타내는 두 항목간의 불일치빈도 b 및 c의 값의 변화에 따라 유사성 측도들의 계산 결과를 나타내면 Table 3.3과 같다.

Table 3.3 Outputs of similarity measures for increasing b and c

a	b	c	d	S_{Yule1}	S_{Digby}	S_{Mich}	S_{Yule2}	S_{Bau}	S_{Disp}
47	3	23	27	0.897	0.798	0.780	0.622	0.480	0.120
46	4	24	26	0.851	0.738	0.737	0.558	0.440	0.110
45	5	25	25	0.800	0.677	0.690	0.500	0.400	0.100
44	6	26	24	0.743	0.615	0.637	0.445	0.360	0.090
43	7	27	23	0.679	0.552	0.581	0.392	0.320	0.080
42	8	28	22	0.610	0.486	0.519	0.340	0.280	0.070
28	22	42	8	-0.610	-0.486	-0.519	-0.340	-0.280	-0.070
27	23	43	7	-0.679	-0.552	-0.581	-0.392	-0.320	-0.080
26	24	44	6	-0.743	-0.615	-0.637	-0.445	-0.360	-0.090
25	25	45	5	-0.800	-0.677	-0.690	-0.500	-0.400	-0.100
24	26	46	4	-0.851	-0.738	-0.737	-0.558	-0.440	-0.110
23	27	47	3	-0.897	-0.798	-0.780	-0.622	-0.480	-0.120
22	28	48	2	-0.937	-0.857	-0.819	-0.694	-0.520	-0.130
21	29	49	1	-0.971	-0.919	-0.853	-0.783	-0.560	-0.140
26	54	4	16	0.316	0.241	0.162	0.156	0.080	0.020
25	55	5	15	0.154	0.116	0.077	0.077	0.040	0.010
24	56	6	14	0.000	0.000	0.000	0.000	0.000	0.000
23	57	7	13	-0.143	-0.108	-0.072	-0.074	-0.040	-0.010
22	58	8	12	-0.275	-0.208	-0.140	-0.145	-0.080	-0.020
21	59	9	11	-0.394	-0.302	-0.205	-0.212	-0.120	-0.030
20	60	10	10	-0.500	-0.390	-0.268	-0.276	-0.160	-0.040
19	61	11	9	-0.594	-0.472	-0.329	-0.335	-0.200	-0.050
18	62	12	8	-0.676	-0.548	-0.389	-0.390	-0.240	-0.060
17	63	13	7	-0.746	-0.619	-0.448	-0.441	-0.280	-0.070
16	64	14	6	-0.806	-0.685	-0.507	-0.487	-0.320	-0.080

이 표에서 보는 바와 같이 불일치빈도 b 또는 c가 증가함에 따라 모든 유사성 측도들이 감소하는 것으로 나타났다. Table 3.2와 마찬가지로 이 표에서도 모든 경우에 S_{Yule1}이 모든 측도들의 상한이 되며, S_{Disp}는 모든 측도들의 하한이 된다. 또한 $|S_{Yule2}| \leq |S_{Digby}|$, $|S_{Bau}| \leq |S_{Yule2}| \leq |S_{Digby}|$의 관계도 성립하였다. 여러 가지 조건에 따라 달라지는 S_{Mich}와 S_{Yule2}, 그리고 S_{Digby}와 S_{Mich}의 대소 관계를 확인하기 위해 먼저 ($a = 49$, $b = 1$, $c = 21$, $d = 29$)의 경우와 ($a = 45$, $b = 5$, $c = 25$, $d = 25$)의 경우를 비교해보면 전자는 $S_{Yule1} = 0.971$, $S_{Digby} = 0.919$, $S_{Mich} = 0.853$, $S_{Yule2} = 0.783$, $S_{Bau} = 0.560$, $S_{Disp} = 0.140$, 그리고 후자는 $S_{Yule1} = 0.800$, $S_{Digby} = 0.677$, $S_{Mich} = 0.690$, $S_{Yule2} = 0.500$, $S_{Bau} = -0.400$, $S_{Disp} = 0.100$으로 나타나서 여기서도 둘 다 $|S_{Disp}| \leq |S_{Mich}| \leq |S_{Yule1}|$ 및 $|S_{Disp}| \leq |S_{Bau}| \leq |S_{Yule2}| \leq |S_{Digby}| \leq |S_{Yule1}|$의 관계가 확인되었다. 반면에 전자의 경우는 [유형 1]의 형태로 나타났고, 후자의 경우에는 [유형 2]의 형태로 나타났다. 이번에는 ($a = 28$, $b = 22$, $c = 42$, $d = 8$)의 경우와 ($a = 23$, $b = 27$, $c = 47$, $d = 3$)의 경우를 비교해볼 때 앞의 예에서와 마찬가지로 두 경우 모두 $|S_{Disp}| \leq |S_{Mich}| \leq |S_{Yule1}|$ 및 $|S_{Disp}| \leq |S_{Bau}| \leq |S_{Yule2}| \leq |S_{Digby}| \leq |S_{Yule1}|$의 관계가 성립하였으며, 전자의 경우는 [유형 2]의 형태로 나타났고, 후자의 경우에는 [유형 1]의 형태로 나타난다는 사실을 확인할 수 있다. 마지막으로 ($a = 21$, $b = 59$, $c = 9$, $d = 11$)의 경우와 ($a = 16$, $b = 64$, $c = 14$, $d = 6$)의 경우를 비교해면 전자의 경우는 [유형 3]의 형태로 나타나고 있고, 후자의 경우에는 [유형 1]의 형태로 나타난다는 사실을 확인할 수 있다.

Warrens (2008)이 기술한 바와 같이 이들 측도들은 각 경계에 있는 측도와는 더욱 더 유사한 값을 가지므로 각 측도의 상한 및 하한은 여러 가지 측도들을 분류하는 도구가 되며, 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

빅 데이터에 내재해 있는 정보를 파악하기 위한 군집 분석 방법은 거리 또는 유사성 측도를 이용하여 각 개체의 유사성을 측정하여 유사도가 높은 대상 집단을 분류하고 군집에 속한 개체들의 유사성과 서로 다른 군집에 속한 개체간의 상이성을 밝혀내는 통계분석 기법이다. 군집분석에서 이용되고 있는 유사성 측도는 데이터의 속성에 따라 여러 가지의 형태로 분류할 수 있는데, 본 논문에서는 주변 확률을 고려하지 않는 확률적 흥미도 측도 계열 유사성 측도들인 Yule I과 II, Michael, Digby, Baulieu, 그리고 Dispersion 측도에 대해 대소 관계를 규명함으로써 각각의 측도에 대한 상한 및 하한의 측도들을 계산하였다. 그 결과를 정리해보면 먼저 $|S_{Yule1}|$은 본 논문에서 고려한 측도들인 $|S_{Yule2}|$, $|S_{Mich}|$, $|S_{Digby}|$, $|S_{Bau}|$, 그리고 $|S_{Disp}|$의 상한이 되고, $|S_{Bau}|$와 $|S_{Yule2}|$는 $|S_{Digby}|$의 하한이 되며, $|S_{Bau}|$는 $|S_{Yule2}|$와 $|S_{Mich}|$의 하한이 된다. 또한 $|S_{Disp}|$는 $|S_{Mich}|$, $|S_{Digby}|$, 그리고 $|S_{Bau}|$의 하한이 된다. 반면에 $|S_{Yule2}|$와 $|S_{Mich}|$, 그리고 $|S_{Mich}|$와 $|S_{Digby}|$는 상황에 따라 대소 관계가 달라지는데, 이들은 $|S_{Yule2}| \leq |S_{Mich}| \leq |S_{Digby}|$, $|S_{Yule2}| \leq |S_{Digby}| \leq |S_{Mich}|$, 그리고 $|S_{Mich}| \leq |S_{Yule2}| \leq |S_{Digby}|$다음과 같이 세 가지 유형의 대소 관계로 요약할 수 있다. 이러한 사실들은 수식의 증명뿐만 아니라 실제 데이터 및 모의실험 데이터에 의해서도 확인되었다.

References

Baulieu, F. B. (1989). A classification of presence/absence based dissimilarity coefficients. *Journal of Classification*, **6**, 233-246.

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.

Gordon, A. D. (1999). *Classification*, Chapman & Hall, London-New York.

Kim, M., Jeon, J., Woo, K. and Kim, M. (2010). A new similarity measure for categorical attribute-based clustering. *Journal of Korean Institute of Information Scientists and Engineers : Databases*, **37**, 71-81.

Lee, J. H. (2013). Big data, data mining and temporary reproduction. *The Journal of Intellectual Property*, **8**, 93-125.

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.

Michael, E. L. (1920). Marine ecology and the coefficient of association. *Journal of Animal Ecology*, **8**, 54-59.

Park, H. C. (2012). Exploration of PIM based similarity measures as association rule thresholds. *Journal of the Korean Data & Information Science Society*, **23**, 1127-1135.

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

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.

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.

Stanfill, C. and Waltz, D. (1986). Toward memory-based reasoning. *Communications of the ACM*, **29**, 1213-1228.

Warrens, M. J. (2008). Bounds of resemblance measures for binary (presence/absence) variables. *Journal of Classification*, **25**, 195-208.

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

Yule, G. U. (1900). On the association of attributes in statistics. *Philosophical Transactions of the Royal Society*, **75**, 257-319.

Yule, G. U. (1912). On the methods of measuring the association between two attributes. *Journal of the Royal Statistical Society*, **75**, 579-652.

Journal of the Korean Data & Information Science Society
2015, **26**(2), 367–376

http://dx.doi.org/10.7465/jkdi.2015.26.2.367
한국데이터정보과학회지

A study on the ordering of PIM family similarity measures without marginal probability

Hee Chang Park[1]

[1]Department of Statistics, Changwon National University

Received 10 February 2015, revised 4 March 2015, accepted 18 March 2015

Abstract

Today, big data has become a hot keyword in that big data may be defined as collection of data sets so huge and complex that it becomes difficult to process by traditional methods. Clustering method is to identify the information in a big database by assigning a set of objects into the clusters so that the objects in the same cluster are more similar to each other clusters. The similarity measures being used in the cluster analysis may be classified into various types depending on the nature of the data. In this paper, we computed upper and lower limits for probability interestingness measure based similarity measures without marginal probability such as Yule I and II, Michael, Digby, Baulieu, and Dispersion measure. And we compared these measures by real data and simulated experiment. By Warrens (2008), Coefficients with the same quantities in the numerator and denominator, that are bounded, and are close to each other in the ordering, are likely to be more similar. Thus, results on bounds provide means of classifying various measures. Also, knowing which coefficients are similar provides insight into the stability of a given algorithm.

Keywords: Big data, cluster analysis, marginal probability, probabilistic interestingness measure, similarity measure.

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