Journal of the Korean Data & Information Science Society
2015, **26**(1), 89–99

http://dx.doi.org/10.7465/jkdi.2015.26.1.89
한국데이터정보과학회지

음의 일치 빈도를 고려한 유사성 측도의 대소 관계 규명에 관한 연구

박희창[1]

[1]창원대학교 통계학과

접수 2014년 11월 26일, 수정 2014년 12월 18일, 게재확정 20140년 12월 26일

요 약

세계 경제 포럼과 대한민국 지식경제부에서 10대 핵심정보기술 가운데 하나로 빅 데이터를 선정한 바 있다. 빅 데이터에 대한 분석은 결국 데이터들이 가지고 있는 속성을 얼마나 효과적으로 분석하느냐가 관건이다. 이를 위한 기법들 중에서 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 있으며, 범주형 데이터에 적용 가능한 측도들은 음의 일치 빈도를 고려한 측도, 음의 일치 빈도를 고려하지 않는 측도, 그리고 주변 확률 분포의 포함 여부에 의한 측도 등으로 구분할 수 있다. 음의 일치 빈도는 동시발생빈도와 더불어 두 항목 간의 관련성에 대한 순방향성을 의미하므로 이를 고려하지 않는 유사성 측도들보다 이를 고려한 유사성 측도들이 좀 더 현실적인 측도라고 할 수 있다. 따라서 본 논문에서는 이분형 데이터에 대해 일반적으로 많이 활용되고 있는 음의 일치 빈도를 고려한 측도들에 대해 대소 관계를 규명함으로써 이들의 상한 및 하한을 설정하는 문제를 고려하였다.

주요용어: 군집 분석, 동시 발생 빈도, 빅 데이터, 유사성 측도, 음의 일치 빈도.

1. 서론

최근 빅 데이터에 대한 관심이 높아지고 있다. 위키 백과사전에 의하면 다양한 종류의 대규모 데이터에 대한 수집 및 분석을 특징으로 하는 빅 데이터 기술의 발전은 다변화된 현대 사회를 더욱 정확하게 예측하여 효율적으로 작동케 하고 과거에는 불가능했던 기술을 실현시키기도 한다. 이러한 연유로 우리나라뿐만 아니라 세계 경제 포럼에서도 10대 핵심정보기술 가운데 하나로 빅 데이터를 선정한 바 있다. 빅 데이터에 대한 분석은 결국 데이터들이 가지고 있는 속성을 얼마나 효과적으로 분석하느냐가 관건이며, 이에 대한 보다 다양한 연구가 필요하다 (Oh 등, 2012).

빅 데이터 분석을 위한 다양한 기법들 중에서 군집 분석 방법은 각 개체 (혹은 대상)의 유사성을 측정하여 유사도가 높은 대상 집단을 분류하고 군집에 속한 개체들의 유사성과 서로 다른 군집에 속한 개체간의 상이성을 밝혀내는 통계분석 기법이다 (Kim, 2009; Jeong, 2005). 군집분석의 방법에는 여러 가지가 있으나 크게 계층적 군집화 방법과 비계층적 군집화 방법으로 나누어진다. 계층적 군집화 방법에는 최단 연결법, 최장 연결법, 평균 연결법, 중심 연결법, 중위수 연결법, 그리고 Ward의 방법 등이 있고, 비계층적 군집화 방법에는 k-평균 군집분석이 대표적인 분석 방법이라고 할 수 있다 (Park, 2009).

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

이러한 군집분석은 아주 방대한 양의 데이터를 대상으로 하고 있으며, 어떤 가정도 없이 자연스럽게 군집을 찾는 기법이므로 정보 검색이나 패턴 인식, 그리고 시장 조사 등 많은 응용 분야에서 널리 사용되고 있다 (Kim 등, 2010). 군집분석에서는 거리 (distance) 또는 유사성 측도 (similarity measure)를 이용하여 동질적인 집단으로 분류하는데 이 때 이용되는 유사성 측도는 데이터의 속성에 따라 여러 가지의 형태로 분류할 수 있으며, 군집분석 및 유사성 측도에 관한 최근 연구로는 Warrens (2008), Choi 등 (2010), Lee와 Kim (2011), Yeo (2011), Park (2011, 2012, 2013), Lim과 Lim (2012), Park과 Kim (2013), Ryu와 Park (2013), Jang 등 (2014), Woo 등 (2014), Cheong과 Oh (2014) 등이 있다.

본 논문에서는 이분형 데이터에 대해 일반적으로 많이 활용되고 있는 음의 일치 빈도를 고려한 측도들에 대해 대소 관계를 규명함으로써 이들의 상한 및 하한을 설정하는 문제를 고려하고자 한다. 이를 위해 Warrens (2008)이 적용한 측도들과 Meyer (2002) 및 Park (2011)에서 논의한 측도들을 동시에 비교한다. 논문의 2절에서는 음의 일치 빈도를 고려한 유사성 측도들을 소개하는 동시에 이들에 대한 상한 및 하한을 계산하는 과정을 기술하며, 3절에서는 실제 예제와 모의실험을 통한 결과를 이용하여 대소 관계를 살펴본 후, 4절에서 결론을 내리고자 한다.

2. 음의 일치 빈도를 고려한 유사성 측도의 대소 관계

음의 일치 빈도는 동시발생빈도와 더불어 두 항목간의 관련성에 대한 순방향성을 의미하므로 이를 고려하지 않는 유사성 측도들보다 이를 고려한 유사성 측도들이 좀 더 현실적인 측도라고 할 수 있다 (Park, 2011). 음의 일치 빈도를 고려한 유사성 측도들을 수식으로 나타내기 위해 다음의 Table 2.1과 같은 분할표를 고려하고자 한다.

Table 2.1 2×2contingency table

		Y		Total
		1	0	
X	1	a	b	$a+b$
	0	c	d	$c+d$
Total		$a+c$	$b+d$	$a+b+c+d$

Meyer (2002)와 Warrens (2008), 그리고 Park (2011)에 의하면 음의 일치 빈도를 고려한 유사성 측도에는 Russel과 Rao 측도, Sokal과 Michener의 단순매칭측도 (simple matching measure), Rogers와 Tanimoto 측도, Sokal과 Sneath 측도, Hamann 측도, 그리고 Baroni-Urbani와 Buser 측도 I, II 등이 있다. 이들을 Table 2.1의 기호를 사용하여 수식으로 나타내면 Table 2.2와 같다.

Table 2.2 Similarity measures with negative matches

similarity measure	formula
Russel & Rao	$S_{RR} = \dfrac{a}{a+b+c+d}$
Sokal & Michener	$S_{SM} = \dfrac{a+d}{a+b+c+d}$
Rogers & Tanimoto	$S_{RT} = \dfrac{a+d}{a+d+2(b+c)}$
Sokal & Sneath	$S_{SS} = \dfrac{2(a+d)}{2(a+d)+b+c}$
Hamann	$S_{HAM} = \dfrac{(a+d)-(b+c)}{a+b+c+d}$
Baroni-Urbani & Buser I	$S_{BUB1} = \dfrac{a+\sqrt{ad}}{a+b+c+\sqrt{ad}}$
Baroni-Urbani & Buser II	$S_{BUB2} = \dfrac{a-b-c+\sqrt{ad}}{a+b+c+\sqrt{ad}}$

위의 식에서 보는 바와 같이 음의 일치 빈도를 고려한 유사성 측도들은 일치 빈도 a 또는 d가 증가하면 이들 측도들은 증가하는 경향을 나타내고, 불일치 빈도 b또는 c가 증가하게 되면 감소하는 경향이 있다. 이들 유사성 측도들에 대한 상한 및 하한을 설정하기 위해 $a+b+c+d$를 n으로 두고 측도 S_{RR}과 다른 측도들과의 대소 관계를 규명하기 위해 먼저 S_{RR}과 S_{SM}을 비교해보면 두 측도의 분모는 동일하고 S_{SM}의 분자가 S_{RR}의 분자보다 크므로 항상 $S_{RR} \leq S_{SM}$이 됨을 쉽게 알 수 있다. S_{RR}과 S_{RT}의 크기를 비교하기 위해 S_{RR}과 S_{RT}의 차이를 구하면 다음과 같다.

$$S_{RR} - S_{RT} = \frac{a(b+c) - nd}{n(n+b+c)}$$

이 식에서 분모는 항상 양의 값을 가지므로 $a(b+c) \leq nd$이면 $S_{RR} \leq S_{RT}$가 되고, 그 반대이면 $S_{RT} \leq S_{RR}$이다. S_{RR}과 S_{SS}의 크기를 비교하기 위해 S_{RR}과 S_{SS}의 차이를 구하면 다음과 같다. 이 식의 분자에서 $a \leq a+2d$이고 동시에 $b+c \leq n$이므로 이들 두 측도의 차이는 음의 값을 가진다. 따라서 항상 $S_{RR} \leq S_{SS}$가 성립한다.

$$S_{RR} - S_{SS} = \frac{a(b+c) - n(a+2d)}{n(n+b+c)}$$

S_{RR}과 S_{HAM}을 비교해볼 때, 이들 두 측도의 분모는 동일하므로 분자만의 차이를 나타내면 $d-(b+c)$이므로 만약 $d \geq b+c$이면 $S_{RR} \leq S_{HAM}$이 되고, $b+c \geq d$이면 $S_{HAM} \leq S_{RR}$이 된다. S_{RR}과 S_{BUB1}의 크기를 비교하기 위해 S_{RR}과 S_{BUB1}의 차이를 구하면 다음의 식과 같이 분자와 분모는 항상 양의 값을 가지므로 항상 $S_{RR} \leq S_{BUB1}$이 성립한다.

$$S_{BUB1} - S_{RR} = \frac{\sqrt{ad}(b+c+d+\sqrt{ad})}{n(a+b+c+\sqrt{ad})}$$

S_{RR}과 S_{BUB2}의 크기를 비교하기 위해 S_{RR}과 S_{BUB2}의 차이를 구하면 다음과 같으므로 분모는 양의 값을 갖게 되어 $b+c+d+\sqrt{ad} \geq n(b+c)$이면 $S_{RR} \leq S_{BUB2}$이 되고, $n(b+c) \geq b+c+d+\sqrt{ad}$이면 $S_{BUB2} \leq S_{RR}$이 된다.

$$S_{RR} - S_{BUB2} = \frac{n(b+c) - (b+c+d+\sqrt{ad})}{n(a+b+c+\sqrt{ad})}$$

다음으로는 측도 S_{SM}과 다른 측도들 간의 관계를 알아보고자 한다. 먼저 S_{SM}과 S_{RT}의 대소 관계를 알아보기 위해 두 측도의 차이를 계산하면 다음과 같다. 따라서 이 식의 분모와 분자 모두 양의 값을 가지므로 항상 $S_{RT} \leq S_{SM}$이 성립한다.

$$S_{SM} - S_{RT} = \frac{(a+d)(b+c)}{n(n+b+c)}$$

S_{SM}과 S_{SS}의 차이를 계산하면 다음 식과 같이 분모와 분자 모두 양의 값을 가지므로 항상 $S_{SM} \leq S_{SS}$가 성립한다.

$$S_{SS} - S_{SM} = \frac{(a+d)(b+c)}{n(n+a+d)}$$

S_{SM}과 S_{HAM}에 대해서는 두 측도의 분모가 동일하고 분자는 S_{SM}이 더 크므로 항상 $S_{HAM} \leq S_{SM}$이 된다. S_{SM}과 S_{BUB1}의 차이를 계산하면 다음과 같으므로 $d \leq a$이면 $S_{SM} \leq S_{BUB1}$이고,

$a \le d$이면 $S_{BUB1} \le S_{SM}$이 된다.

$$S_{BUB1} - S_{SM} = \frac{(b+c)\sqrt{d}(\sqrt{a}-\sqrt{d})}{n(a+b+c+\sqrt{ad})}$$

S_{SM}과 S_{BUB2}의 차이를 계산하면 다음 식과 같이 항상 $\sqrt{a} \le n\sqrt{d}$이 성립하므로 $S_{BUB2} \le S_{SM}$이 된다.

$$S_{BUB2} - S_{SM} = \frac{\sqrt{d}(\sqrt{a}-n\sqrt{d})}{n(a+b+c+\sqrt{ad})}$$

측도 S_{RT}과 다른 측도들 간의 관계를 알아보기 위해 먼저 S_{RT}와 S_{SS}의 차이를 계산하면 다음과 같으므로 항상 $S_{RT} \le S_{SS}$가 성립한다.

$$S_{SS} - S_{RT} = \frac{3(a+d)(b+c)}{(n+a+d)(n+b+c)}$$

S_{RT}와 S_{HAM}의 차이를 계산하면 다음과 같게 되어 항상 $S_{HAM} \le S_{RT}$가 성립한다.

$$S_{RT} - S_{HAM} = \frac{2(b+c)^2}{n(n+b+c)}$$

S_{RT}와 S_{BUB1}의 차이를 계산하면 다음과 같다.

$$S_{BUB} - S_{RT} = \frac{(b+c)[\sqrt{a}+\sqrt{d})^2 - 2d]}{(n+b+c)(a+b+c+\sqrt{ad})}$$

따라서 $(\sqrt{a}+\sqrt{d})^2 \le 2d$이면 $S_{BUB1} \le S_{RT}$이고, 그 반대이면 $S_{RT} \le S_{BUB1}$가 된다.

S_{RT}와 S_{BUB2}의 차이를 계산하면 다음과 같으므로 $\sqrt{ad} \le b+c+d$이면 $S_{BUB2} \le S_{RT}$이고, 그 반대이면 $S_{RT} \le S_{BUB2}$가 된다.

$$S_{BUB2} - S_{RT} = \frac{2(b+c)[\sqrt{ad} - (b+c+d)]}{(n+b+c)(a+b+c+\sqrt{ad})}$$

이제 측도 S_{SS}와 다른 측도들 간의 관계를 알아보기 위해 먼저 S_{SS}와 S_{HAM}의 차이를 계산하면 다음과 같고, 이 식에서 $a+d \le n$이고 $a+d-b-c \le n$이므로 분자는 항상 양의 값을 가지게 되어 $S_{HAM} \le S_{SS}$가 성립한다.

$$S_{SS} - S_{HAM} = \frac{n^2 - (a+d)[a+d-(b+c)]}{n(n+a+d)}$$

S_{SS}와 S_{BUB1}의 차이는 다음과 같이 계산된다.

$$S_{SS} - S_{BUB1} = \frac{(b+c)(a+2d-\sqrt{ad})}{(n+a+d)(a+b+c+\sqrt{ad})}$$

기하평균 (G)과 산술평균 (A)의 관계식 $G(a,d) \le A(a,d)$로부터 $\sqrt{ad} \le (a+d)/2$가 되고, $(a+d)/2 \le a+2d$이므로 항상 $S_{BUB1} \le S_{SS}$이 된다. 그리고 Table 2.2로부터 $S_{BUB2} \le S_{BUB1}$임을 알 수 있고, $S_{BUB1} \le S_{SS}$이므로 항상 $S_{BUB2} \le S_{SS}$가 성립한다.

S_{HAM}과 다른 측도들과의 대소 관계를 알아보기 위해 먼저 S_{HAM}과 S_{BUB1}의 차이를 계산하면 다음과 같이 정리되므로 따라서 항상 $S_{HAM} \leq S_{BUB1}$이 성립한다.

$$S_{BUB1} - S_{HAM} = \frac{b+c}{n}$$

S_{HAM}과 S_{BUB2}의 차이를 계산하면 다음의 식과 같이 나타낼 수 있으므로 $d \leq a$이면 $S_{HAM} \leq S_{BUB2}$이고, $a \leq d$이면 $S_{BUB2} \leq S_{HAM}$이 된다.

$$S_{BUB2} - S_{HAM} = \frac{2(b+c)\sqrt{d}(\sqrt{a}-\sqrt{d})}{n(a+b+c+\sqrt{ad})}$$

마지막으로 S_{BUB1}과 S_{BUB2}에 대해서는 Table 2.2로부터 항상 $S_{BUB2} \leq S_{BUB1}$이 성립함을 알 수 있다.

이러한 측도들 간의 관계로부터 S_{SS}는 본 논문에서 고려하는 모든 유사성 측도들의 상한이 됨을 알 수 있고, 항상 $S_{HAM} \leq S_{RT} \leq S_{SM} \leq S_{SS}$, $S_{RR} \leq S_{SM}$, $S_{BUB2} \leq S_{SM}$, $S_{HAM} \leq S_{BUB1}$, $S_{BUB2} \leq S_{BUB1}$, $S_{RR} \leq S_{BUB1}$이 성립한다. 여러 가지 조건에 따른 측도들 간의 관계를 기술하면 Table 2.3과 같다.

Table 2.3 Upper and lower bounds by conditions

bounds	condition
$S_{SM} \leq S_{BUB1}$	$d \leq a$
$S_{HAM} \leq S_{BUB2}$	$d \leq a$
$S_{HAM} \leq S_{RR}$	$d \leq b+c$
$S_{RR} \leq S_{RT}$	$a(b+c) \leq nd$
$S_{BUB2} \leq S_{RT}$	$\sqrt{ad} \leq b+c+d$
$S_{BUB1} \leq S_{RT}$	$(\sqrt{a}+\sqrt{d})^2 \leq 2d$
$S_{BUB2} \leq S_{RR}$	$b+c+d+\sqrt{ad} \leq n(b+c)$
$S_{RR} \leq S_{SM} \leq S_{SS}$	none
$S_{RR} \leq S_{BUB1} \leq S_{SS}$	none
$S_{BUB2} \leq S_{SM} \leq S_{SS}$	none
$S_{HAM} \leq S_{BUB1} \leq S_{SS}$	none
$S_{BUB2} \leq S_{BUB1} \leq S_{SS}$	none
$S_{HAM} \leq S_{RT} \leq S_{SM} \leq S_{SS}$	none

Table 2.3으로부터 모든 측도들의 대소 관계를 규명할 수 있는 동시에 이들 측도간의 상한 및 하한을 구할 수 있다. 예를 들어 $a(b+c) \leq nd$, $d \leq b+c$, $d \leq a$, 그리고 $\sqrt{ad} \leq b+c+d$이면 $S_{HAM} \leq S_{RR} \leq S_{RT} \leq S_{BUB2} \leq S_{BUB1} \leq S_{SM} \leq S_{SS}$가 된다.

Warrens (2008)이 기술한 바와 같이 이들 측도들은 각 경계에 있는 측도와는 더욱 더 유사한 값을 가지므로 각 측도의 경계 값은 여러 가지 측도들을 분류하는 도구가 되며, 실제 값의 관점에서 각 측도들이 유사하다는 것을 알게 되면 주어진 알고리즘의 안정화에 도움이 될 수 있다.

3. 적용 예제

이 절에서는 음의 일치 빈도를 고려하지 않은 유사성 측도들의 실제적인 변화 양성을 고찰하기 위해 2003년 경남사회지표조사 자료를 이용하였다. 이 조사는 조사대상 모집단을 경상남도 내 총 가구 수로 하였고, 이에 대한 표본 수는 총 가구 수의 10%정도인 10,000가구로 하였다.

Table 3.1 Outputs of similarity measures without negative matches of survey data

items	a	b	c	d	S_{BUB2}	S_{HAM}	S_{RR}	S_{RT}	S_{BUB1}	S_{SM}	S_{SS}
$X1$: $X5$	592	3,480	558	2,611	-0.375	-0.115	0.082	0.284	0.312	0.442	0.613
$X1$: $X6$	2,086	3,322	1,864	2,721	-0.074	-0.038	0.209	0.317	0.463	0.481	0.650
$X1$: $X7$	221	5,185	419	4,165	-0.652	-0.122	0.022	0.281	0.174	0.439	0.610
$X1$: $X8$	1,450	3,954	1,313	3,270	-0.184	-0.055	0.145	0.309	0.408	0.473	0.642
$X1$: $X9$	530	4,870	545	4,039	-0.462	-0.085	0.053	0.297	0.269	0.458	0.628
$X1$: $X10$	1,074	4,331	899	3,686	-0.261	-0.047	0.108	0.313	0.369	0.476	0.645
$X1$: $X11$	349	5,042	300	4,277	-0.546	-0.072	0.035	0.302	0.227	0.464	0.634
$X2$: $X5$	160	673	990	5,418	-0.208	0.541	0.022	0.626	0.396	0.770	0.870
$X2$: $X6$	469	733	3,481	5,310	-0.346	0.157	0.047	0.407	0.327	0.578	0.733
$X2$: $X7$	113	1,089	527	8,261	-0.199	0.676	0.011	0.722	0.400	0.838	0.912
$X2$: $X8$	362	839	2,401	6,385	-0.265	0.351	0.036	0.510	0.367	0.676	0.806
$X2$: $X9$	146	1,056	929	7,853	-0.240	0.602	0.015	0.668	0.380	0.801	0.890
$X2$: $X10$	248	954	1,725	7,063	-0.261	0.464	0.025	0.577	0.370	0.732	0.845
$X2$: $X11$	88	1,109	561	8,210	-0.281	0.665	0.009	0.713	0.360	0.832	0.909
$X3$: $X5$	210	1,263	940	4,828	-0.288	0.392	0.029	0.533	0.356	0.696	0.821
$X3$: $X6$	791	1,255	3,159	4,788	-0.234	0.117	0.079	0.387	0.383	0.558	0.717
$X3$: $X7$	134	1,911	506	7,439	-0.362	0.516	0.013	0.610	0.319	0.758	0.862
$X3$: $X8$	594	1,452	2,169	5,772	-0.194	0.275	0.059	0.468	0.403	0.637	0.779
$X3$: $X9$	263	1,782	812	7,127	-0.228	0.480	0.026	0.588	0.386	0.740	0.851
$X3$: $X10$	397	1,649	1,576	6,368	-0.238	0.354	0.040	0.512	0.381	0.677	0.808
$X3$: $X11$	138	1,906	511	7,413	-0.355	0.515	0.014	0.610	0.322	0.758	0.862
$X4$: $X5$	43	187	1,107	5,904	-0.406	0.643	0.006	0.697	0.297	0.821	0.902
$X4$: $X6$	115	204	3,835	5,839	-0.624	0.192	0.012	0.424	0.188	0.596	0.747
$X4$: $X7$	25	294	615	9,056	-0.290	0.818	0.003	0.833	0.355	0.909	0.952
$X4$: $X8$	95	223	2,668	7,001	-0.521	0.421	0.010	0.551	0.240	0.711	0.831
$X4$: $X9$	38	281	1,037	8,628	-0.367	0.736	0.004	0.767	0.317	0.868	0.929
$X4$: $X10$	66	253	1,907	7,764	-0.468	0.568	0.007	0.644	0.266	0.784	0.879
$X4$: $X11$	25	294	624	9,025	-0.295	0.816	0.003	0.831	0.353	0.908	0.952

자료의 구조는 크게 일반사항 (인구통계학적 문항)과 도민의식조사부문으로 나누어져 있으며, 도민의식조사부문 중에서 환경정책 시급과제로 주민의 환경의식 개혁 및 강화 ($X1$), 환경사범에 대한 벌칙 강화 ($X2$), 수질개선 등 환경보호시설 확충 ($X3$), 민간 환경단체의 조직 및 기능 강화 ($X4$)의 4개 문항에 대해 자녀의 환경오염 저감 행동 ($X5$), 1회용품 사용 유무 ($X6$), 산림의 중요성 체감 ($X7$), 사회복지정책 만족도 ($X8$), 환경의식 개혁 필요성 ($X9$), 최우선 행정 추진 분야로서의 환경보존 ($X10$), 차선 행정 추진 분야로서의 환경보존 ($X11$)의 7개 문항에 대해 유사성 측도를 계산하여 Table 3.1에 제시하였다. 이 표에서 $X3$과 $X10$의 문항을 예로 들면 a는 환경정책 시급과제로 수질개선 등 환경보호시설 확충이라고 응답하고 최우선 행정 추진 분야로 환경보존을 응답한 사람들의 수, b는 환경정책 시급과제로 환경보호시설 확충이라고 응답하고 최우선 행정 추진 분야로 환경보존 이외의 문항에 응답한 사람들의 수, c는 환경정책 시급과제로 환경보호시설 확충 이외의 문항에 응답하고 최우선 행정 추진 분야로 환경보존을 응답한 사람들의 수, 그리고 d는 환경정책 시급과제로 환경보호시설 확충 이외의 문항에 응답하고 최우선 행정 추진 분야로 환경보존 이외의 문항에 응답한 사람들의 수를 의미한다. 이 표에서 보는 바와 같이 모든 경우에 S_{SS}는 측도들의 상한이 되며, 항상 $S_{RR} \leq S_{SM}$, $S_{RR} \leq S_{BUB1}$, $S_{BUB2} \leq S_{SM}$, $S_{HAM} \leq S_{BUB1}$, $S_{BUB2} \leq S_{BUB1}$, 그리고 $S_{HAM} \leq S_{RT} \leq S_{SM}$이 성립한다. 이

외의 경우에는 Table 2.3에서 제시한 바와 같이 여러 가지 조건에 따라 대소 관계가 달라지는 것을 알 수 있다. 이를 좀 더 구체적으로 알아보기 위해 먼저 S_{HAM}과 S_{RR}을 비교해보면 [$X1$: $X7$]에서 보는 바와 같이 $d \leq b+c$이면 $S_{HAM} \leq S_{RR}$이 성립하고 $d \geq b+c$이면 [$X3$: $X8$]에서 보는 바와 같이 $S_{HAM} \geq S_{RR}$이 성립함을 알 수 있다. 또한 $(\sqrt{a}+\sqrt{d})^2 \leq 2d$인 [$X4$: $X5$] 경우에는 $S_{BUB1} \leq S_{RT}$가 되고, $(\sqrt{a}+\sqrt{d})^2 \geq 2d$인 [$X1$: $X8$]의 경우에는 $S_{BUB1} \geq S_{RT}$가 된다.

이와 같이 실제 예제를 통해 확인하는 데에는 한계가 있으므로 Park (2011)에서와 같이 여러 가지 모의실험 자료를 이용하여 좀 더 구체적으로 음의 일치 빈도를 고려한 유사성 측도들의 대소 관계를 살펴보는 것이 필요하다. 이를 위해 먼저 항목 X와 Y에 대한 음의 일치 빈도 $d = n(X=0,\ Y=0)$의 변화에 따라 여러 가지 유사성 측도들의 변화 양상을 살펴보기 위해 Table 3.2를 활용하고자 한다.

Table 3.2 Simulation data(1)

		Y		Total
		1	0	
X	1	$20+d$	$30-d$	50
	0	$50-d$	d	50
Total		70	30	100

Table 3.2에서 보는 바와 같이 전체 트랜잭션의 수를 100명, 항목 X의 발생빈도는 50명, 그리고 항목 Y의 발생빈도를 70명으로 하였다. 항목 집합 X와 Y의 음의 일치 빈도는 d명으로 하였으며, d가 취할 수 있는 범위는 $0 \leq d \leq 30$이다. 이 표를 이용하여 d의 변화에 따라 음의 일치 빈도를 고려한 유사성 측도들을 계산한 결과를 나타내면 Table 3.3과 같다. 이 표에서 $a = n(X=1,\ Y=1)$, $b = n(X=1,\ Y=0)$, 그리고 $c = n(X=0,\ Y=1)$이다. 이 표에서 $a = n(X=1,\ Y=1)$, $b = n(X=1,\ Y=0)$, 그리고 $c = n(X=0,\ Y=1)$이다.

Table 3.3 Comparison of similarity measures by simulation data(1)

a	b	c	d	S_{HAM}	S_{BUB2}	S_{RT}	S_{RR}	S_{SM}	S_{BUB1}	S_{SS3}
30	20	40	10	-0.200	-0.118	0.250	0.300	0.400	0.441	0.571
31	19	39	11	-0.160	-0.079	0.266	0.310	0.420	0.460	0.592
32	18	38	12	-0.120	-0.041	0.282	0.320	0.440	0.480	0.611
33	17	37	13	-0.080	-0.003	0.299	0.330	0.460	0.499	0.630
34	16	36	14	-0.040	0.035	0.316	0.340	0.480	0.518	0.649
35	15	35	15	0.000	0.073	0.333	0.350	0.500	0.537	0.667
36	14	34	16	0.040	0.111	0.351	0.360	0.520	0.556	0.684
37	13	33	17	0.080	0.149	0.370	0.370	0.540	0.574	0.701
38	12	32	18	0.120	0.186	0.389	0.380	0.560	0.593	0.718
39	11	31	19	0.160	0.224	0.408	0.390	0.580	0.612	0.734
40	10	30	20	0.200	0.261	0.429	0.400	0.600	0.631	0.750
41	9	29	21	0.240	0.299	0.449	0.410	0.620	0.649	0.765
42	8	28	22	0.280	0.336	0.471	0.420	0.640	0.668	0.780
43	7	27	23	0.320	0.373	0.493	0.430	0.660	0.686	0.795
44	6	26	24	0.360	0.410	0.515	0.440	0.680	0.705	0.810
45	5	25	25	0.400	0.447	0.538	0.450	0.700	0.724	0.824
46	4	24	26	0.440	0.484	0.563	0.460	0.720	0.742	0.837
47	3	23	27	0.480	0.521	0.587	0.470	0.740	0.761	0.851
48	2	22	28	0.520	0.558	0.613	0.480	0.760	0.779	0.864
49	1	21	29	0.560	0.595	0.639	0.490	0.780	0.798	0.876
50	0	20	30	0.600	0.632	0.667	0.500	0.800	0.816	0.889

이 표에서 보는 바와 같이 a와 d의 값이 증가함에 따라 본 논문에서 고려하는 모든 유사성 측도들이 증가하는 것으로 나타났다. S_{SS}는 가장 크게 나타나서 모든 측도들의 상한값이 되며, $S_{RR} \leq S_{SM}$,

$S_{RR} \leq S_{BUB1}$, $S_{BUB2} \leq S_{SM}$, $S_{HAM} \leq S_{BUB1}$, $S_{BUB2} \leq S_{BUB1}$, 그리고 $S_{HAM} \leq S_{RT} \leq S_{SM}$의 관계가 성립함을 알 수 있다. 또한 이 표에서는 항상 $d \leq a$이므로 $S_{HAM} \leq S_{BUB2}$및 $S_{SM} \leq S_{BUB1}$의 관계가 성립한다. 특히 S_{HAM}과 S_{BUB2}인 경우에는 절대값의 크기가 a와 d의 크기에 따라 대소 관계가 결정된다. 이외의 측도들에 대해서는 조건에 따라 대소 관계가 여러 가지 형태로 나타났다. 이를 좀 더 구체적으로 확인하기 위해 먼저 S_{HAM}과 S_{RR}을 비교해보면 $d \leq b + c$이 되는 ($a = 47$, $b = 3$, $c = 23$, $d = 27$)의 경우까지는 $S_{HAM} \leq S_{RR}$이 되고 그 이외에는 부등호가 반대가 된다. S_{RR}과 S_{RT}의 관계를 살펴보면 $a(b + c) \geq nd$이 되는 ($a = 37$, $b = 13$, $c = 33$, $d = 17$)의 경우까지는 $S_{RR} \geq S_{RT}$이 성립하고, 그 이하에서는 $S_{RR} \leq S_{RT}$가 된다. 모든 경우에 $\sqrt{ad} \leq b + c + d$, $(\sqrt{a} + \sqrt{d})^2 \leq 2d$, 그리고 $b + c + d + \sqrt{ad} \leq n(b + c)$이므로 이 표에서는 항상 $S_{BUB2} \leq S_{RT}$, $S_{BUB1} \leq S_{RT}$, 그리고 $S_{BUB2} \leq S_{RR}$가 성립한다.

이번에는 두 항목간의 불일치빈도 b의 값의 변화에 따라 정확도들의 변화하는 양상을 파악하기 위해 Park (2013)에서와 같이 Table 3.4와 같은 분할표를 이용하고자 한다. 이 표에서 b가 취할 수 있는 정수 값의 범위는 $0 \leq b \leq 30$이다.

Table 3.4 Simulation data(2)

		Y		Total
		1	0	
X	1	$30 - b$	b	30
	0	$b + 20$	$50 - b$	70
Total		50	50	100

이 표로부터 불일치빈도 b값의 변화에 따른 유사성 측도들을 계산하면 다음 Table 3.5와 같은 결과를 얻을 수 있다.

Table 3.5 Comparison of similarity measures by simulation data(2)

a	b	c	d	S_{RR}	S_{BUB2}	S_{HAM}	S_{RT}	S_{BUB1}	S_{SM}	S_{SS}
25	5	25	45	0.250	0.322	0.400	0.538	0.661	0.700	0.824
24	6	26	44	0.240	0.277	0.360	0.515	0.638	0.680	0.810
23	7	27	43	0.230	0.231	0.320	0.493	0.616	0.660	0.795
22	8	28	42	0.220	0.185	0.280	0.471	0.593	0.640	0.780
21	9	29	41	0.210	0.140	0.240	0.449	0.570	0.620	0.765
20	10	30	40	0.200	0.094	0.200	0.429	0.547	0.600	0.750
19	11	31	39	0.190	0.048	0.160	0.408	0.524	0.580	0.734
18	12	32	38	0.180	0.002	0.120	0.389	0.501	0.560	0.718
17	13	33	37	0.170	-0.045	0.080	0.370	0.478	0.540	0.701
16	14	34	36	0.160	-0.091	0.040	0.351	0.455	0.520	0.684
15	15	35	35	0.150	-0.137	0.000	0.333	0.431	0.500	0.667
14	16	36	34	0.140	-0.184	-0.040	0.316	0.408	0.480	0.649
13	17	37	33	0.130	-0.231	-0.080	0.299	0.384	0.460	0.630
12	18	38	32	0.120	-0.279	-0.120	0.282	0.361	0.440	0.611
11	19	39	31	0.110	-0.326	-0.160	0.266	0.337	0.420	0.592
10	20	40	30	0.100	-0.374	-0.200	0.250	0.313	0.400	0.571
9	21	41	29	0.090	-0.423	-0.240	0.235	0.289	0.380	0.551
8	22	42	28	0.080	-0.472	-0.280	0.220	0.264	0.360	0.529
7	23	43	27	0.070	-0.522	-0.320	0.205	0.239	0.340	0.507
6	24	44	26	0.060	-0.572	-0.360	0.190	0.214	0.320	0.485
5	25	45	25	0.050	-0.625	-0.400	0.176	0.188	0.300	0.462
4	26	46	24	0.040	-0.678	-0.440	0.163	0.161	0.280	0.438
3	27	47	23	0.030	-0.735	-0.480	0.149	0.133	0.260	0.413
2	28	48	22	0.020	-0.796	-0.520	0.136	0.102	0.240	0.387
1	29	49	21	0.010	-0.866	-0.560	0.124	0.067	0.220	0.361
0	30	50	20	0.000	-1.000	-0.600	0.111	0.000	0.200	0.333

이 표에서 보는 바와 같이 불일치빈도 b가 증가함에 따라 모든 유사성 측도들이 감소하는 것으로 나타났다. Table 3.3과 마찬가지로 이 표에서도 S_{SS}는 모든 측도들의 상한값이 되었으며, $S_{RR} \leq S_{SM}$, $S_{RR} \leq S_{BUB1}$, $S_{BUB2} \leq S_{SM}$, $S_{HAM} \leq S_{BUB1}$, $S_{BUB2} \leq S_{BUB1}$, 그리고 $S_{HAM} \leq S_{RT} \leq S_{SM}$으로 나타났다. 또한 이 표에서는 모든 사례에서 $a \leq d$이므로 $S_{BUB2} \leq S_{HAM}$및 $S_{BUB1} \leq S_{SM}$의 관계가 성립하는 동시에 $\sqrt{ad} \leq b + c + d$이 되므로 $S_{BUB2} \leq S_{RT}$가 되며, $a(b + c) \leq nd$이므로 $S_{RR} \leq S_{RT}$이 성립한다. 이외의 측도들에 대해서는 대소 관계가 조건에 따라 여러 가지 형태로 나타났다. 이를 좀 더 구체적으로 확인하기 위해 먼저 S_{HAM}과 S_{RR}을 비교해보면 $b + c \leq d$이 되는 ($a = 20$, $b = 10$, $c = 30$, $d = 40$)의 경우까지는 $S_{RR} \leq S_{HAM}$이 되고 그 이외에는 부등호의 방향이 바뀐다. S_{BUB1}과 S_{RT}의 관계를 살펴보면 $(\sqrt{a} + \sqrt{d})^2 \geq 2d$이 되는 ($a = 5$, $b = 25$, $c = 45$, $d = 25$)의 경우까지는 $S_{RT} \leq S_{BUB1}$이 되고 b가 25 보다 크면 부등호가 반대로 된다. S_{BUB2}와 S_{RR}의 경우에는 ($a = 23$, $b = 7$, $c = 27$, $d = 43$)의 경우까지는 $b + c + d + \sqrt{ad} \leq n(b + c)$가 되어 $S_{RR} \leq S_{BUB2}$이 되고, $b + c + d + \sqrt{ad} \leq n(b + c)$이면 $S_{BUB2} \leq S_{RR}$이 된다.

4. 결론

본 논문에서는 일반적으로 많이 활용되고 있는 음의 일치 빈도를 고려한 측도인 Russel과 Rao 측도 S_{RR}, Sokal과 Michener의 단순매칭측도 S_{SM}, Rogers와 Tanimoto 측도 S_{RT}, Sokal과 Sneath 측도 S_{SS}, Hamann 측도 S_{HAM}, 그리고 Baroni-Urbani와 Buser 측도 S_{BUB1} 및 S_{BUB2} 등에 대해 대소 관계를 규명함으로써 이들의 상한 및 하한을 설정하는 문제를 고려하였다. 그 결과, S_{SS}는 본 논문에서 고려하는 모든 유사성 측도들의 상한이 됨을 알 수 있었고, 항상 $S_{HAM} \leq S_{RT} \leq S_{SM} \leq S_{SS}$, $S_{RR} \leq S_{SM}$, $S_{BUB2} \leq S_{SM}$, $S_{HAM} \leq S_{BUB1}$, $S_{BUB2} \leq S_{BUB1}$, $S_{RR} \leq S_{BUB1}$이 된다는 사실을 확인하였다. 또한 여러 가지 조건에 따른 측도들 간의 관계를 살펴보았는데, $d \leq a$이면 $S_{SM} \leq S_{BUB1}$과 $S_{HAM} \leq S_{BUB2}$가 성립하였으며, $d \leq b+c$이면 $S_{HAM} \leq S_{RR}$가 되고, $a(b+c) \leq nd$이면 $S_{RR} \leq S_{RT}$, $\sqrt{ad} \leq b + c + d$이면 $S_{BUB2} \leq S_{RT}$, $(\sqrt{a} + \sqrt{d})^2 \leq 2d$이면 $S_{BUB1} \leq S_{RT}$, 그리고 $b + c + d + \sqrt{ad} \leq n(b + c)$이면 $S_{BUB2} \leq S_{RR}$이 된다는 사실을 수식의 증명뿐만 아니라 실제 데이터 및 모의실험 데이터에 의해서도 확인하였다.

References

Cheong, D. and Oh, K. J. (2014). Using cluster analysis and genetic algorithm to develop portfolio investment strategy based on investor information. *Journal of the Korean Data & Information Science Society*, **25**, 107-117.

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.

Jang, H., Kim, K. K. and Kang, C. (2014). Comparison of clustering methods for categorical data. *Journal of the Korean Data Analysis Society,* **16**, 2439-2445.

Jeong, K. M. (2005). A note on Bayesian information criterion in model-based clustering. *Journal of the Korean Data Analysis Society*, **7**, 1517-1529.

Kim, D. (2009). On the Silhouette plot in cluster analysis. *Journal of the Korean Data Analysis Society*, **11**, 2955-2964.

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, 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.

Meyer A. (2002) *Comparison of similarity coefficients used in cluster analysis with dominant markers data*, MSc Thesis, Universidade de Sao Paulo, Piracicaba.

Oh, S. M., Song, J. M. and Kim, C. S. (2012). Clustering analysis using the influence of attributes in categorical data analysis. *Journal of the Korean Institute of Information Scientists and Engineers*, **18**, 790-793.

Park, H. C. (2009). *An introduction to statistical database*, Changwon National University Press, Changwon.

Park, H. C. (2011). Association rule thresholds of similarity measures considering negative co-occurrence frequencies. *Journal of the Korean Data & Information Science Society*, **22**, 1113-1122.

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. (2013). Proposition of causal association rule thresholds. *Journal of the Korean Data & Information Science Society,* **24**, 1189-1197.

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.

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

Woo, S. Y., Lee, J. W. and Jhun, M. (2014). Microarray data analysis using relative hierarchical clustering. *Journal of the Korean Data & Information Science Society*, **25**, 999-1009.

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**(1), 89–99

http://dx.doi.org/10.7465/jkdi.2015.26.1.89
한국데이터정보과학회지

A study on the ordering of similarity measures with negative matches

Hee Chang Park[1]

[1]Department of Statistics, Changwon National University

Received 26 November 2014, revised 18 December 2014, accepted 26 December 20140

Abstract

The World Economic Forum and the Korean Ministry of Knowledge Economy have selected big data as one of the top 10 in core information technology. The key of big data is to analyze effectively the properties that do have data. Clustering analysis method of big data techniques is a method of assigning a set of objects into the clusters so that the objects in the same cluster are more similar to each other clusters. 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 studied upper and lower bounds for binary similarity measures with negative matches such as Russel and Rao measure, simple matching measure by Sokal and Michener, Rogers and Tanimoto measure, Sokal and Sneath measure, Hamann measure, and Baroni-Urbani and Buser mesures I, II. And the comparative studies with these measures were shown by real data and simulated experiment.

Keywords: Big data, cluster analysis, co-occurrence frequency, negative matches, similarity measures.

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