

유사도 통합에 관한 연구

김선경¹⁾, 박지수^{**}, 손진곤²⁾

*한국방송통신대학교 정보과학과

**전주대학교 컴퓨터공학과

skkim@wizbase.co.kr, jisupark@jj.ac.kr, jgshon@knou.ac.kr

A Study on Integrating Similarities

Sunkyung Kim*, Ji Su Park**, Jin Gon Shon*

*Dept. of Computer Science, Korea National Open University

**Dept. of Computer Science and Engineering, Jeonju University

요 약

유사도는 두 객체의 비슷한 정도를 실수로 나타낸 것이며 반대 개념인 다른 정도를 나타내는 것을 거리라 한다. 실세계에서 정확히 같은 것은 존재하기 힘들기 때문에 많은 응용 분야에서 유사도나 거리를 이용한다. 거리 중 대표적인 것으로 유클리드 공간에서 두 점 사이의 직선거리이다. 이 거리를 유클리드 거리라고 한다. 코사인 유사도는 벡터 공간에서 두 벡터 사이각의 코사인 값이다. 이외에도 용도에 따라 다양한 거리 또는 유사도가 연구되고 있다. 수학적으로 유사도는 이변수 함수로 나타낸다. 앞선 연구에서 민코프스키는 맨하탄 거리, 유클리드 거리 등을 매개변수 p 를 이용하여 하나의 식으로 통합하였다. 이러한 유사도 통합은 유사도에 대한 새로운 통찰력을 제공하고 또 다른 응용을 제공한다. 본 논문은 기존 유사도의 의미를 개관하고 추가적인 매개변수를 도입하여 민코프스키 거리와 코사인 유사도를 통합한 식을 제시한다.

1. 서론

어떤 대상 간 대소 여부 혹은 정확한 일치 여부를 판단하는 비교 연산은 사람보다 컴퓨터가 빠르고 정확하게 수행한다. 그러나 실세계의 많은 문제는 어제 본 사람이 지금 보고 있는 사람과 같은 사람인가를 판단해야 하는 것처럼 비슷한 정도를 판단해야 한다. 그리고 사람은 실생활에서 비슷한 정도를 큰 노력 없이 계속해서 판단하고 있다. 컴퓨터가 이러한 판단을 수행하기 위해서 유사도(Similarity)를 구하는 연산을 한다. 유사도의 반대 개념인 다른 정도는 거리(Distance)라고 불리는 개념을 이용한다[1,2]. 이런 유사성 또는 비유사성을 측정하는 연산을 할 때 컴퓨터는 단순 비교 연산보다 많은 비용이 소모한다.

유사도는 얼굴이나 인식이나 표정 인식을 통한 감정 인식과 같은 형상 인식, 단백질의 유사성 파악을 통한 신약 개발 등 실세계의 많은 문제를 해결하기 위해 이용되고 있다[3-6]. 2,000년 이전 유클리드 거리부터 시작해서 현대의 코사인 유사도[7], 자

카드 유사도[8], 민코프스키 거리[9] 등 많은 유사도가 연구되어 오고 있다. 이런 유사도들은 비슷한 형태로 묶여지며 비슷한 유사도는 민코프스키거리와 같이 매개변수를 이용하여 통합된다[10]. 유사도 통합을 통하여 새로운 통찰력과 응용을 제공한다. 본 논문은 코사인 유사도를 거리 개념으로 변경한 후 민코프스키 거리와 통합한 거리를 제시한다.

본 논문은 제2장에서 거리 및 유사도에 대하여 개관한다. 제3장에서는 코사인 유사도와 동일한 특성을 가지는 거리를 제시한다. 제4장에서는 점의 호 거리라는 거리를 제시하고 유클리드 거리와 관계를 밝힌다. 제5장에서 코사인 유사도와 민코프스키 거리를 통합하는 거리를 제시한다. 제6장은 본 연구에 대한 결론을 짓는다.

2. 거리/유사도 개관

거리는 두 개의 서로 다른 대상이 얼마나 떨어져 있는지를 나타내는 실수 값이다. 거리는 거리 함수를 이용하여 계산된다. 거리 함수의 정의역은 거리를 측정하고자 하는 대상 집합의 카티션 곱이며 공역은 실수이다[11]. 거리 함수는 아래와 같은 조건을 가져야 한다.

1) 한국방송통신대학교 대학원 재학생

2) 교신저자

X 는 집합이다.

거리 함수 $d: X \times X \rightarrow \mathbb{R}$ 일 때 $x, y, z \in X$ 이다.

조건 1. $d(x, y) \geq 0$

조건 2. $d(x, y) = 0$ 이면 $x = y$ 이다.

조건 3. $d(x, y) = d(y, x)$

조건 4. $d(x, y) \leq d(x, z) + d(z, y)$

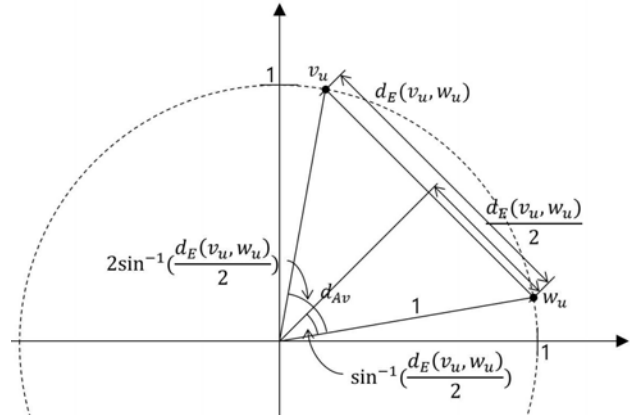
조건 1은 거리가 음의 수가 나올 수 없다는 조건이다. 거리는 떨어져 있는 정도라는 개념이므로 음의 값이 나올 수가 없다. 조건 2는 거리가 0이면 비교 대상이 되었던 두 대상은 같은 것이어야 한다는 것이다. 이 또한 우리가 일반적으로 생각하는 거리라는 개념과 일치한다. 조건 3은 x 에서 y 로의 거리와 y 에서 x 로의 거리는 같다는 것이다. 조건 4는 다른 곳으로 돌아가는 거리가 더 짧은 것은 존재하지 않는다는 것이다. 조건 3, 4번 역시 우리가 일반적으로 인식하는 거리와 일치하는 개념이다. 쉽게 생각할 수 있는 거리는 유클리드 공간에서 두 점 사이의 직선거리이며 이것을 유클리드 거리라고 한다. 유클리드 거리는 점의 각 성분 차이를 제곱하여 합한 후 제곱근을 하여 계산하며 조건 1부터 조건 4까지 만족함을 수학적으로 쉽게 증명 가능하다[12].

거리의 반대 개념으로 유사도가 있다. 유사도는 두 대상이 얼마나 비슷한지를 나타내는 수치를 계산하는 함수를 말한다. 유사도 함수의 정의역은 대상 집합의 카티션 곱이며 공역은 실수이다[11]. 거리와 반대로 유사도는 값이 클수록 두 대상은 가까이 있다. 대부분의 유사도 함수는 두 대상이 같을 때 유사도 값이 1이 된다. 또한 거리와 유사도는 간단한 식을 통하여 쉽게 서로 변환된다. 많은 분야에서 이용되고 있는 유사도는 두 벡터 사이각의 코사인 값을 취한 코사인 유사도이다. 두 벡터가 같은 방향을 가리킬 때 사이각은 0라디안이며 코사인 값은 1이고 점점 다른 방향을 가리킬 때 사이각은 점점 커지며 코사인 값은 점점 작아져서 사이각이 π 라디안일 때 -1 까지 된다. 즉 코사인 유사도는 두 벡터가 동일한 방향일 때 완전히 동일하다는 의미에서 1의 값을 수직일 때는 같지 않다는 의미에서 0을 서로 완전히 반대 방향일 때는 -1 을 값을 가진다.

3. 코사인 유사도의 특징을 가지는 거리

코사인 유사도는 두 벡터 사이각의 코사인 값이다. 벡터 공간에서 벡터의 스칼라적을 두 벡터 크기의 곱으로 나누어준 것으로 계산한다[12].

벡터의 길이는 코사인 유사도에 영향을 주지 않으며 방향만을 영향을 미친다. 코사인 유사도를 반



(그림 1) 벡터의 호 거리

대 개념인 거리로 정의하기 위하여 코사인 유사도와 동일하게 벡터의 길이가 거리에 영향이 미치지 않도록 단위 벡터로 변환하겠다. 그러면 단위 벡터는 단위원에 위치하며 단위원에서 두 벡터를 잇는 짧은 호의 거리를 ‘벡터의 호 거리’라 하겠다. 벡터의 호 거리는 코사인 유사도와 동일하게 벡터의 방향에만 영향을 받는 거리이며 라디안 각이므로 쉽게 코사인 유사도로 변환이 가능한 값이다. 두 벡터 v, w 의 단위 벡터를 v_u, w_u 라고 할 때 호의 길이는 식(1)과 같다. (그림 1)을 보면 두 단위 벡터와 원점을 잇는 이등변 삼각형을 이용하게 식(1)이 유도된다.

$$d_{Av} = 2\sin^{-1}\left(\frac{d_E(v_u, w_u)}{2}\right) \quad (1)$$

여기서 d_E 는 유클리드 거리이다.

코사인 유사도를 거리로 변환하는 경우 유클리드 거리와 삼각함수로 유도됨을 확인하였다.

4. 유클리드 거리와 호 거리

좌표 공간에서 두 점을 지나는 반지름이 r 인 원의 호 거리를 ‘점의 호 거리’라 정의하겠다. 두 점 v, w 에 대한 점의 호 거리는 식(2)와 같다. 식(1)과 유사한 방식으로 식(2)도 두 점과 반지름이 r 인 원의 중심을 잇는 이등변 삼각형을 이용하여 식을 유도된다. 단, 원의 지름 $2r$ 이 두 점의 거리보다 짧으면 원이 두 점을 지나지 못한다. 이 경우 거리를 무한대로 정의하였다.

$$d_{Ap} = \begin{cases} 2r\sin^{-1}\left(\frac{d_E(v, w)}{2r}\right), & d_E(v, w) \leq 2r \\ \infty, & d_E(v, w) > 2r \end{cases} \quad (2)$$

여기서 d_E 는 유클리드 거리이다.

식(2)에서 반지름인 r 이 무한대로 갈 때를 생각해 보자. d_{Ap} 를 극한을 취하면 식(3)에서 보듯이 d_{Ap} 는 유클리드 거리가 된다. r 이 무한대로 가는 극한 식으로 쓴 후 $\frac{1}{2r}$ 을 t 로 치환하였다. r 이 무한대로 갈 때 $\frac{1}{2r}$ 은 0으로 수렴하므로 t 가 0으로 수렴하는 식으로 다시 썼다. 분수식에서 0으로 수렴할 때 분자와 분모를 미분한 식도 동일하다는 로피탈의 정리를 미분한 후 식을 정리하면 유클리드 거리가 된다. 따라서 r 값을 매개변수로 할 경우 점의 호 거리는 유클리드 거리와 통합된다.

$$\begin{aligned} \lim_{r \rightarrow \infty} d_{Ap} &= \lim_{r \rightarrow \infty} 2r \sin^{-1} \left(\frac{d_E(v,w)}{2r} \right) \\ t &= \frac{1}{2r}, \left(\frac{1}{2r} \text{를 } t \text{로 치환} \right) \\ \lim_{t \rightarrow 0} d_{Ap} &= \lim_{t \rightarrow 0} \frac{\sin^{-1}(d_E(v,w)t)}{t} \\ &= \lim_{t \rightarrow 0} \frac{(\sin^{-1}(d_E(v,w)t))'}{t'} \\ &= \lim_{t \rightarrow 0} \frac{d_E(v,w)}{\sqrt{1-(d_E(v,w)t)^2}} \\ &= \lim_{t \rightarrow 0} \frac{1}{1} \\ &= d_E(v,w) \end{aligned} \tag{3}$$

또한 유클리드 거리(d_E)를 매개변수 p 에 따라 맨하탄, 유클리드 거리 등을 표현하는 민코프스키 거리 ($\sqrt[p]{\sum_{i=1}^n |v_i - w_i|^p}$)로 치환할 경우 점의 호 거리와 민코프스키 거리와 점의 호 거리가 통합이 된다.

5. 유사도 통합

민코프스키 거리는 매개변수 p 에 따라서 1인 경우 맨하탄 거리, 2인 경우 유클리드 거리, 무한대인

경우 체비쇼프 거리가 된다[9]. 또한 점의 호 거리는 반지름을 1로 하고 벡터를 단위 벡터로 변환할 경우 제3장에 정의한 벡터의 호 거리가 되며 벡터의 호 거리는 코사인 유사도와 같은 특성을 가지는 거리임을 확인했다. 따라서 제4장에서 정의한 점의 호 거리에 매개변수를 적절히 취하였을 경우 코사인 유사도와 민코프스키 거리가 통합이 된다.

$$d_A = \begin{cases} 2r \sin^{-1} \left(\frac{d_M}{2r} \right), & d_M \geq 2r \\ \infty, & d_M < 2r \end{cases} \tag{4}$$

여기서 $d_M = \sqrt[p]{\sum_{i=1}^n |av_i - bw_i|^p}$

식(4)는 매개변수 p, r, a, b 를 적용하여 매개변수에 따라 6가지의 거리를 나타내는 식이다.<표 1>은 각 매개변수 값에 따른 점의 호 거리가 나타내는 거리를 보여준다.

통합된 식에서 r 값을 조정함에 따라 단위원상의 호의 거리가 되거나 직선거리가 된다. p 값은 민코프스키 거리에서 사용된 방식과 동일하게 사용된다. a, b 값은 두 벡터의 길이가 거리에 영향을 주는 정도를 조정된다. a, b 값이 1이면 벡터의 길이가 그대로 영향을 주며 벡터의 길이 역수를 적용하면 벡터의 길이 영향이 없어진다. 응용할 할 때 벡터마다 길이의 영향을 다르게 주는 거리를 필요하면 a, b 값을 조정한다.

6. 결론

코사인 유사도를 단위 벡터가 단위원에서 호를 지나는 거리로 변환하여 코사인 유사도처럼 벡터의 크기에 영향을 받지 않고 방향만 영향을 받는 벡터의 호 거리를 제시하였다. 두 점을 지나는 반지름이 r 인 원의 호 거리를 점의 호 거리로 정의하고 r 이 무한대로 갈 때 점의 호 거리는 유클리드 거리임을 보였다. 유클리드 거리를 민코프스키 거리로 치환하

<표 1> 매개 변수에 따른 점의 호 거리가 나타내는 거리

거리	r	p	a	b
점의 호 거리	$(0, \infty]$	2	1	1
벡터의 호 거리(d_{Av}) 코사인 유사도($\cos(d_{Av})$)	1	2	$\frac{1}{ v }$	$\frac{1}{ w }$
맨하탄 거리	∞	1	1	1
유클리드 거리	∞	2	1	1
체비쇼프 거리	∞	∞	1	1
민코프스키 거리	∞	$(0, \infty]$	1	1

고 매개변수 2개를 추가하여 벡터의 호 거리와 점의 호 거리를 통합하였다. 매개변수 p, r, a, b 를 이용하여 통합된 점의 호 거리가 매개변수에 따라 코사인 유사도와 민코프스키 거리를 나타냄을 보였다.

제시된 점의 호 거리는 <표 1>에서 제시된 매개변수 이외의 값을 취하여 좀 더 다양한 거리 함수가 만들어진다. 이렇게 만들어진 거리 함수는 데이터베이스 분야에서 유사도 질의에서 응용된다. 그리고 인공지능의 패턴인식 분야에서 최근접이웃 탐색 등에서 응용된다. 이후 연구에서 매개변수 값이 <표 1>에 제시된 이외의 값을 적용한 유사도가 실제 사례에서 어떤 의미를 가지고 기여 가능한 응용이 존재하는지 찾아보겠다. 그리고 거리가 아닌 유사도 형태로의 통합식을 추후 연구를 통하여 제시하겠다.

참고문헌

[1] Similarity measure, https://en.wikipedia.org/wiki/Similarity_measure, [Acceded by 2020.09.27]

[2] Metric (mathematics), [https://en.wikipedia.org/wiki/Metric_\(mathematics\)](https://en.wikipedia.org/wiki/Metric_(mathematics)), [Acceded by 2020.09.27]

[3] Richard O. Duda , Peter E. Hart , David G. Stork, “Pattern Classification” 2nd Edition, Wiley-Interscience, 2000

[4] Dongshun Cui, Guang-Bin Huang, Tianchi Liu, “ELM based smile detection using Distance Vector”, Pattern Recognition 79 pp. 356-369, 2018

[5] P.F. Felzenszwalb, R.B. Girshick, D. McAllester, D. Ramanan, “Object detection with discriminatively trained part-based models”, IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI) 32 (9) pp. 1627 - 1645, 2010

[6] 이지영, 최지원, 신재민, “컴퓨터를 이용한 신약 개발, 향후 기술 전망”, 정보과학회지 제31권 제8호, pp. 35-54, 2013

[7] Cosine similarity, https://en.wikipedia.org/wiki/Cosine_similarity, [Acceded by 2020.09.27]

[8] Jaccard index, https://en.wikipedia.org/wiki/Jaccard_index, [Acceded by 2020.09.27]

[9] Minkowski distance, https://en.wikipedia.org/wiki/Minkowski_distance,

[Acceded by 2020.09.27]

[10] Sung-Hyuk Cha, “Comprehensive Survey on Distance/Similarity Measures between Probability Density Functions”, International Journal of Mathematical models and Methods in Applied Sciences, 1 (4), pp. 300-307, 2007

[11] Michel Marie Deza, Elena Deza “Encyclopedia of Distance” 4th Ed. Springer, 2006

[12] David Poole, “Linear Algebra: A Modern introduction” 4th Ed, Cengage Learning, 2016