Journal of the Korean Data & Information Science Society
2012, **23**(2), 227–234

한국데이터정보과학회지

의학진단에 이용되는 해밍 거리의 특성 탐색[†]

안정용[1]

[1]전북대학교 통계학과

접수 2011년 12월 23일, 수정 2012년 1월 13일, 게재확정 2012년 1월 25일

요 약

의학진단을 위해 여러 증상과 질병 사이의 거리를 이용하는 연구가 많이 진행되고 있다. 그러나 거리들이 비슷한 값을 가지는 경우가 많이 발생하며, 이들 거리의 차이값은 정규분포 또는 카이제곱분포 등과 같은 일반적인 통계분포를 따르지 않는다. 본 연구에서는 의학진단에 사용되는 해밍 거리들의 차이값에 대한 분포적 특성에 대해 살펴보고, 이 차이값의 유의성 검정에 대해 탐색해보고자 한다.

주요용어: 구간값 퍼지 데이터, 의학 진단, 해밍 거리.

1. 서론

의학 분야에서 퍼지 데이터의 활용은 Zadeh (1969)의 제안으로부터 시작되어, 질병과 증상 사이의 관계를 모델링한 Sanchez (1976, 1979)의 연구를 계기로 많은 발전을 이루게 되었다. 그 후 Atanassov (1986)에 의해 직관적 퍼지 집합 (intuitionistic fuzzy sets)의 개념이 소개되어 의학진단 과정을 모델링하기 위한 여러 방법들에 활용되고 있으며 (Adlassnig, 1986; De 등, 2001; Innocent와 John, 2004), 최근 여러 학문 분야의 기법들을 의학분야에 적용하기 위한 연구들이 많이 이루어지고 있다 (Park과 Kim, 2010; Jo와 Baik, 2010; Jo, 2011; Joung과 Chung, 2011).

퍼지 데이터를 의학진단에 이용하는 가장 일반적인 방법은 퍼지 데이터에 기반하여 질병과 증상 사이의 관계를 정의하고, 이들 관계에 max-min-max 규칙을 적용하여 환자의 질병을 판단하는 방법이다. 그러나 이 방법은 최대값 또는 최소값의 극단적인 정보만을 이용하기 때문에 이들 데이터를 제외한 다른 데이터들이 가지고 있는 정보의 손실이 발생하는 단점을 가지고 있다. 예를 들어, 환자가 가지고 있는 증상들의 정도가 0.7, 0.4, 0.6 이라 할 때 max-min-max 규칙을 이용하면 진단 척도의 값이 0.4가 될 수 있다. 그러나 일반적으로 환자가 갖는 증상들은 어떤 질병과 강하게 연결되어 있기 때문에 진단 척도의 값은 0.4보다는 큰 값을 갖는 것이 타당하다.

이러한 문제점을 해결하기 위한 방법 중 하나는 데이터들 사이의 유사성 또는 거리를 이용하는 것이다. 퍼지 집합들 사이의 거리척도는 의학진단 분야는 물론 사회과학, 경제학, 공학 등 다른 분야에서도 많이 활용되고 있으며 (Guo 등, 2010; Liang과 Shi, 2003; Wang과 Xin, 2005; Zeng 등, 2009), 구간값을 갖는 퍼지 데이터의 거리척도를 의학진단에 이용한 연구는 Szmidt 등 (2001)에 의해 시작되었다. 그러나 현재까지 진행된 대부분의 연구들은 질병과 증상사이의 아주 단순한 관계만을 이용한다. 또한, 현실세계에서 발생하는 질병의 대부분은 매우 다양한 증상들을 수반하고, 이러한 증상들은 여러 질병과 관

[†] 이 논문은 2011년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (2011-0008954).
[1] (561-756) 전라북도 전주시 덕진구 백제대로 567, 전북대학교 통계학과 (응용통계연구소), 교수.
E-mail: jyahn@jbnu.ac.kr

계되어 있다. 따라서 환자가 가지고 있는 증상들의 정보를 서로 관련되어 있는 질병들 단위로 집계하는 과정이 필수적이지만 이러한 사항을 고려하는 연구는 거의 이루어지지 않고 있다.

본 연구에서는 구간값 퍼지 데이터를 이용한 의학진단에서 진단 척도로 흔히 활용되는 해밍 거리 (Hamming distances)의 특성에 관해 살펴보고자 한다. 특히, 질병을 진단하기 위하여 해밍 거리의 차이를 이용하므로 거리의 차이에 대한 유의성을 판단하는 기준을 제시해보고자 한다. 2장에서는 의학진단을 위해 본 연구에서 사용하는 구간값 퍼지 데이터의 형태에 대해 소개하고, 해밍 거리를 이용하는 의학진단 과정을 기술한다. 3장에서는 해밍 거리의 분포적 특성에 대해 살펴보고, 거리의 차이에 대한 유의성에 대해 탐색한다.

2. 의학진단 방법

2.1. 데이터 형태

의학진단을 위한 데이터 개발에서 가장 먼저 고려해야 할 사항은 진단하고자 하는 질병의 범위를 정하는 것이다. 질병의 범위가 결정되면 각 질병과 관련되어 나타날 수 있는 증상 목록을 작성한다. 물론 이들 증상들은 하나의 질병에서만 나타나는 것은 아니며, 각 질병과 크고 작은 관련성을 가지고 있다. 이러한 관련성에 따라 구간값 퍼지 정도를 부여한다.

본 연구에서 이용되는 데이터는 선행연구 Kim 등 (2007), Ahn 등 (2011)을 통해 개발되었으며, 질병의 범위를 두통에서 가장 흔하게 나타나는 편두통 (migraine), 긴장형두통 (tension headache), 군발성두통 (cluster headache)으로 한정하였다. 편두통과의 관련성이 큰 증상 23개 ($M1 \sim M23$), 긴장형두통과의 관련성이 큰 증상 17개 ($T1 \sim T17$), 군발성두통과의 관련성이 큰 증상 15개 ($C1 \sim C15$) 목록을 작성하였으며, 각각의 증상들에는 3가지 두통과의 관련 정도에 따라 표 2.1과 같이 확신 및 불확신 퍼지 정도가 구간값으로 부여되어 있다.

표 2.1 데이터 형태

증상번호	증상	구간값 퍼지 정도					
		편두통		긴장형두통		군발성두통	
		확신 정도	불확신 정도	확신 정도	불확신 정도	확신 정도	불확신 정도
M1	Positive family history ...	[0.5, 0.6]	[0.2, 0.3]	[0.2, 0.3]	[0.4, 0.6]	[0.2, 0.3]	[0.5, 0.6]
M2	At least five attacks ...	[0.7, 0.8]	[0.1, 0.2]	[0.1, 0.2]	[0.6, 0.7]	[0.1, 0.2]	[0.6, 0.7]
:	:	:	:	:	:	:	:
T1	At least 10 previous ...	[0.3, 0.4]	[0.5, 0.6]	[0.7, 0.8]	[0.1, 0.2]	[0.1, 0.2]	[0.6, 0.7]
:	:	:	:	:	:	:	:
C15	Composite ...	[0.2, 0.3]	[0.5, 0.7]	[0.1, 0.3]	[0.6, 0.7]	[0.7, 0.8]	[0.1, 0.2]

2.2. 진단 과정

데이터의 개발과 더불어 가장 중요한 과정은 의학진단 방법을 개발하는 것이다. 본 연구에서 이용하는 의학 진단 방법은 선행연구 Ahn 등 (2011)을 통해 개발되었으며 여기에서는 그 절차에 대해서만 간단히 소개한다. 진단 방법은 다음과 같이 4단계로 이루어져 있다.

- 단계 1 : 환자가 갖는 증상에 대한 고유한 정도 (patient's degrees; 이하 '고유 정도')와 일반적인 확신 정도 (confirmability degrees; 이하 '일반 정도')를 수집한다. 일반 정도는 증상과 질병과의 관련정도를 나타내는 것으로 표 2.1에 정리되어 있다. 고유 정도는 환자와 증상과의 관계를 나타내는 값으로 의사에 의해 할당된다. 다시 말하면, 일반 정도는 증상과 질병과의 일반적인 관계를 나타내는 값인 반면, 고유 정도는 특정 환자와 증상사이의 고유한 관계 (같은 증상이라도 환자에 따라 심한 정도가 다를 것이다)를 나타내는 값이다.

표 2.2 환자 P_1의 고유 정도

증상	M8	M9	M14	M18	T4	T11	T13	C2	C4
M	[0.6, 0.7]	[0.6, 0.7]	[0.7, 0.8]	[0.5, 0.6]	[0.4, 0.5]	[0.4, 0.5]	[0.5, 0.6]	[0.5, 0.6]	[0.6, 0.7]
N	[0.1, 0.2]	[0.1, 0.2]	[0.0, 0.1]	[0.2, 0.3]	[0.2, 0.3]	[0.2, 0.3]	[0.1, 0.3]	[0.1, 0.2]	[0.2, 0.3]

- 단계 2 : 고유 정도와 일반 정도 각각의 IIFWAA (interval-valued intuitionistic fuzzy weighted arithmetic average)를 계산한다. IIFWAA는 여러 증상들에 대한 집계정보이며, 정의 2.1의 연산자를 이용하여 계산한다.
- 단계 3 : 단계 2에서 계산된 IIFWAA 사이의 해밍 거리를 계산한다. 해밍 거리의 계산은 Park 등 (2008)에서 제시된 식을 수정한 정의 2.2를 이용한다.
- 단계 4 : 단계 3에서 계산된 해밍 거리에 기반하여 환자의 질병을 판단한다. 일반적으로 가장 작은 해밍 거리를 갖는 질병을 환자의 우선적인 질병으로 진단한다.

정의 2.1 (IIFWAA 연산자) $A = \{< x_i, M_A(x_i), N_A(x_i) >| \ i = 1, 2, ..., n\}$를 구간값을 갖는 퍼지 데이터의 집합이라 하면 IIFWAA는 다음과 같이 정의된다.

$$IIFWAA(A) = ([1 - \Pi_{i=1}^{n}(1 - M_{AL}(x_i))^{\omega_i}, \ 1 - \Pi_{i=1}^{n}(1 - M_{AU}(x_i))^{\omega_i}],$$

$$[\Pi_{i=1}^{n}(N_{AL}(x_i))^{\omega_i}, \ \ \Pi_{i=1}^{n}(N_{AU}(x_i))^{\omega_i}])$$

여기에서 n은 퍼지 데이터의 수, $M_A(x_i)$는 $[M_{AL}(x_i), M_{AU}(x_i)]$으로 질병의 확신정도에 대한 구간값이며, $N_A(x_i)$는 $[N_{AL}(x_i), N_{AU}(x_i)]$으로 불확신 정도를 나타내는 구간값이다. 즉, $M_{AL}(x_i)$과 $N_{AL}(x_i)$은 각 구간의 하한값, $M_{AU}(x_i)$과 $N_{AU}(x_i)$은 각 구간의 상한값이며, x_i는 구간값 $M_A(x_i)$와 $N_A(x_i)$를 갖는 i번째 퍼지 데이터를 나타낸다. IIFWAA(A)는 퍼지 데이터 집합 A를 이용하여 계산한, 질병의 확신정도 구간값과 불확신정도 구간값을 구성요소로 가지고 있다. 또한, $\omega = (\omega_1, \omega_2, ..., \omega_n)^T$는 A의 가중치 벡터이며, $\omega_i > 0$, $\sum_{i=1}^{n} \omega_i = 1$이다. 본 연구에서는 $\omega = (1/n, 1/n, ..., 1/n)$을 이용한다.

정의 2.2 (해밍 거리) $A = \{< x_i, M_A(x_i), N_A(x_i) >| \ i = 1, 2, ..., n\}$, $B = \{< x_i, M_B(x_i), N_B(x_i) >| \ i = 1, 2, ..., n\}$를 구간값을 갖는 퍼지 데이터의 집합이라 하면 해밍 거리는 다음과 같이 정의된다.

$$l(A, B) = (1/4n) \sum [\ |M_{AL}(x_i) - M_{BL}(x_i)| + |M_{AU}(x_i) - M_{BU}(x_i)|$$

$$+|N_{AL}(x_i) - N_{BL}(x_i)| + |N_{AU}(x_i) - N_{BU}(x_i)|$$

$$+|H_{AL}(x_i) - H_{BL}(x_i)| + |H_{AU}(x_i) - H_{BU}(x_i)| \ ]$$

여기에서 H는 불확정 값으로 $H_{AL}(x_i) = 1 - (M_{AU}(x_i) + N_{AU}(x_i))$, $H_{AU}(x_i) = 1 - (M_{AL}(x_i) + N_{AL}(x_i))$이다.

예제 2.1 어떤 환자가 편두통의 (M8, M9, M14, M18), 긴장형두통의 (T4, T11, T13), 군발성두통의 (C2, C4)와 같은 증상들을 가지고 있다고 하자. 이 환자의 질병 진단 과정은 다음과 같다.

- 단계 1 : 먼저, 고유 정도와 일반 정도를 수집한다. 표 2.2는 의사에 의해 할당된 고유 정도, 표 2.3은 표 2.1에 제시된 일반 정도이다.
- 단계 2 : 표 2.2와 표 2.3의 데이터에 IIFWAA 연산자를 적용하여 표 2.4, 표 2.5와 같이 IIFWAA를 계산한다.

표 2.3 환자 P_1의 일반 정도

	편두통		긴장형두통		군발성두통	
증상	확신 정도	불확신 정도	확신 정도	불확신 정도	확신 정도	불확신 정도
M8	[0.6, 0.7]	[0.1, 0.2]	[0.2, 0.3]	[0.5, 0.6]	[0.4, 0.5]	[0.4, 0.5]
M9	[0.6, 0.7]	[0.1, 0.2]	[0.3, 0.4]	[0.4, 0.6]	[0.3, 0.4]	[0.4, 0.5]
M14	[0.5, 0.6]	[0.2, 0.3]	[0.1, 0.2]	[0.6, 0.7]	[0.2, 0.3]	[0.6, 0.7]
M18	[0.6, 0.7]	[0.2, 0.3]	[0.2, 0.4]	[0.4, 0.6]	[0.4, 0.6]	[0.1, 0.2]
T4	[0.4, 0.5]	[0.3, 0.5]	[0.5, 0.6]	[0.2, 0.3]	[0.3, 0.4]	[0.3, 0.4]
T11	[0.2, 0.3]	[0.6, 0.7]	[0.6, 0.7]	[0.1, 0.2]	[0.0, 0.1]	[0.7, 0.8]
T13	[0.1, 0.3]	[0.5, 0.6]	[0.6, 0.7]	[0.1, 0.2]	[0.0, 0.1]	[0.6, 0.8]
C2	[0.4, 0.5]	[0.3, 0.5]	[0.2, 0.3]	[0.4, 0.5]	[0.6, 0.7]	[0.2, 0.3]
C4	[0.5, 0.6]	[0.2, 0.3]	[0.1, 0.2]	[0.6, 0.7]	[0.6, 0.7]	[0.1, 0.2]

- 단계 3 : 표 2.4와 표 2.5로부터 표 2.6의 해밍 거리를 계산한다.
- 단계 4 : 표 2.6의 해밍 거리에 기반하여 환자의 질병을 예비 진단 (preliminary diagnosis) 한다. 이 예제의 환자는 편두통으로부터 가장 많은 고통을 겪고 있는 것으로 판단되며 추가적인 검사를 통해 질병을 확진한다.

표 2.4 환자 P_1의 고유 정도에 대한 IIFWAA

	편두통 증상	긴장형두통 증상	군발성두통 증상
P_1	([0.61, 0.71], [0.00, 0.19])	([0.44, 0.54], [0.16, 0.30])	([0.56, 0.65], [0.14, 0.25])

표 2.5 환자 P_1의 일반 정도에 대한 IIFWAA

	편두통	긴장형두통	군발성두통
편두통 증상	([0.58, 0.68], [0.14, 0.24])	([0.20, 0.33], [0.47, 0.62])	([0.33, 0.46], [0.31, 0.43])
긴장형두통 증상	([0.24, 0.37], [0.45, 0.59])	([0.57, 0.67], [0.13, 0.23])	([0.11, 0.21], [0.50, 0.63])
군발성두통 증상	([0.45, 0.55], [0.24, 0.39])	([0.15, 0.25], [0.49, 0.59])	([0.60, 0.70], [0.14, 0.24])

표 2.6 환자 P_1의 각 질병에 대한 해밍 거리

	편두통	긴장형두통	군발성두통
P_1	0.172	0.329	0.222

3. 해밍 거리의 특성

해밍 거리를 이용하여 질병을 진단하는 일반적인 방법은 가장 작은 거리를 갖는 항목을 환자의 질병으로 진단하는 것이다. 그러나 거리들의 차이가 매우 작게 나타나는 경우가 있을 수 있다. 예를 들어, 표 2.6에서 편두통과 군발성두통의 차이는 0.05이다. 이러한 경우 환자의 질병을 편두통 하나로만 진단할 것인지, 아니면 편두통과 군발성두통을 동시에 가지는 것으로 진단할 것인지가 문제가 된다. 이러한 문제를 해결하기 위해서는 이 차이값들의 분포적 특성에 대한 탐색이 필요하다. 해밍 거리의 차이값들에 대한 통계적 분포를 파악하면 이에 대한 확률을 이용할 수 있고 진단 결과를 세밀하게 분석할 수 있기 때문이다.

해밍 거리는 컴퓨터 통신 등에서 문자열의 전송 도중 몇개의 글자에서 오류가 났는지를 측정하기 위한 척도로 Hamming (1950)에 의해 소개되었으며, 암호학, 최적화 문제, 정보이론 등의 분야에서 많이 활용되고 있다. 의학진단분야에서는 Szmidt 등 (2001)에 의해 이용이 시작되었으나 해밍 거리에 대한 통계적 특성을 탐구하는 연구는 거의 이루어지지 않고 있다. 물론 Kammerdiner 등 (2010)에서 해밍 거리의 통계적 분포에 관한 연구가 일부 이루어졌지만 이 연구는 퍼지 데이터에 관련된 것이 아니기 때문에 본 연구와 같이 퍼지 데이터를 활용하는 의학진단 분야에서 직접 이용할 수 없다.

해밍 거리의 차이값에 대한 특성을 탐색하기 위해 예제 2.1과 같이 가상 환자에 대한 시뮬레이션을 실시하였다. 예제 2.1은 9개의 증상을 갖는 환자의 각 질병에 대한 해밍 거리를 계산해놓은 것이다. 본 연구에서는 환자가 갖는 증상의 수를 3개에서 21개까지 변화시켜 가면서 각각의 경우에 환자 10,000명에 대한 해밍 거리를 구하였다. 예를 들어, 환자가 갖는 증상의 수가 3개일때의 시뮬레이션 과정은 다음과 같다. 먼저, 55개의 증상 중 3개의 증상을 임의로 배정하고, 환자가 갖는 증상에 대한 고유 정도와 일반 정도를 수집한다. 일반 정도는 표 2.1의 값을 이용하고, 고유 정도는 전문가의 의견을 참고하여 0.3 이상의 값을 임의로 할당한다. 두번째 과정에서는 정의 2.1을 이용하여 IIFWAA를 계산한다. 마지막으로 정의 2.2를 이용하여 해밍 거리를 계산한다. 이러한 과정을 10,000번 반복하여 환자 10,000명에 대한 해밍 거리를 얻는다.

그림 3.1은 해밍 거리들의 산점도이다. 대각선 아랫부분의 산점도는 증상의 수가 9개일때의 것이고, 대각선 윗부분은 증상의 수가 17개일때의 산점도이다. 환자가 갖는 증상의 수가 17개 정도로 많을 경우에 각 거리들간에 특별한 패턴은 보이지 않는다. 그러나 증상의 수가 적은 경우 편두통과 긴장형두통, 긴장형두통과 군발성두통은 3개의 군집형태를 보이며, 편두통과 군발성두통은 비교적 강한 정상관 관계(상관계수는 약 0.47)를 보인다. 따라서 편두통과 군발성두통은 동시에 나타나거나 동시에 나타나지 않을 가능성이 크다고 할 수 있다.

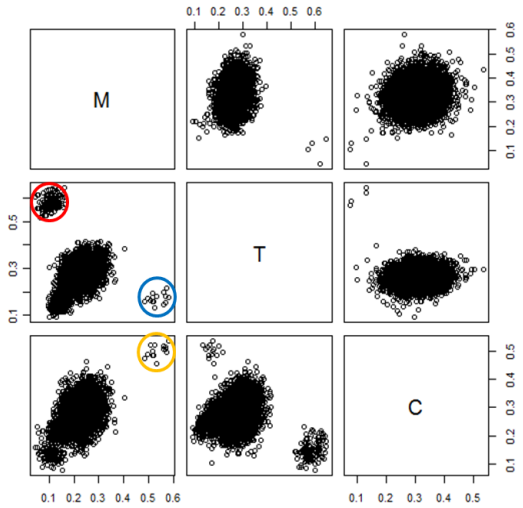

그림 3.1 해밍 거리들 사이의 관계 (증상이 9개, 17개인 경우)

그림 3.2는 환자가 갖는 증상의 수가 9개일때의 편두통과 긴장형두통의 차이값, 긴장형두통과 군집성두통의 차이값, 군발성두통과 편두통의 차이값에 대한 분포이다. 그림에서 보는 바와 같이 편두통과

긴장형두통의 차이값, 긴장형두통과 군집성두통의 차이값은 봉우리가 2개인 형태를 가지고 있다. 그림 3.2의 첫번째 히스토그램의 작은 봉우리는 그림 3.1에서 빨간색과 파란색 원으로 표시된 데이터 값들이다. 빨간색으로 표시된 값들은 긴장형두통 증상이 나타나지 않았을때, 파란색으로 표시된 값들은 편두통 증상이 나타나지 않았을때의 값들이다. 노란색으로 표시된 값들은 편두통 증상이 나타나지 않으면서 긴장형두통 증상이 상대적으로 많이 나타날 때의 값들이다.

그림에서 보는 바와 같이 정규분포 또는 카이제곱분포 등과 같은 일반적인 분포를 따르지 않는 형태이며, 이러한 현상은 증상의 수가 작을때 (11이하인 경우) 심하게 나타난다. 따라서 시뮬레이션 자료를 이용하여 해밍 거리의 차이에 대한 유의성 검정을 하는 것이 타당함을 알 수 있다. 표 3.7은 유의수준 5%에서 해밍 거리의 차이에 대한 임계값 (critical value)을 증상의 수에 따라 정리해 놓은 것이다. 즉, 이 임계값들은 10,000개의 시뮬레이션 값들 중 제 95백분위수에 해당되는 값이다. 증상의 수가 9일때 편두통과 긴장형두통의 차이에 대한 임계값은 약 0.144이며, 편두통과 군발성두통의 차이에 대한 임계값은 약 0.137이다. 따라서 거리가 가장 작은 값을 기준으로 생각하면 표 2.6의 경우에는 환자가 편두통과 군발성두통을 같이 가지고 있는 것으로 진단하는 것이 타당하다고 하겠다.

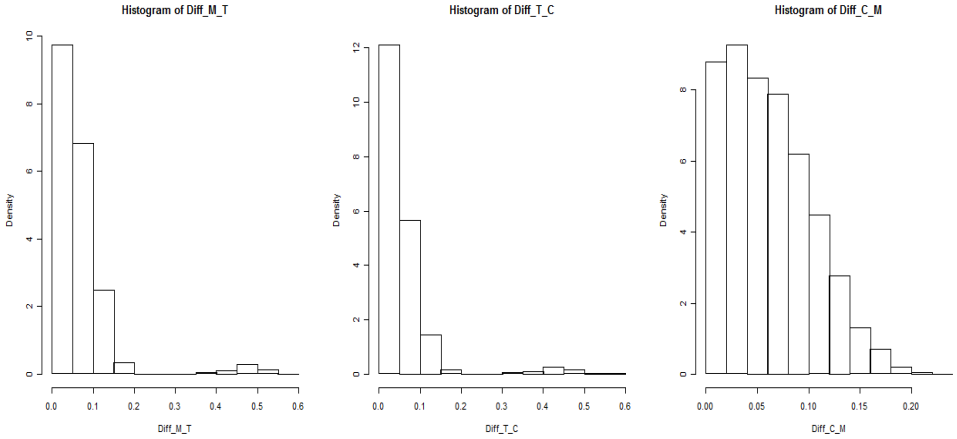

그림 3.2 해밍 거리 차이값의 분포

표 3.1 임계값

증상의 수	3	5	7	9	11	13	15	17	19	21
편두통과 긴장형두통의 차이	0.503	0.482	0.429	0.144	0.128	0.119	0.114	0.110	0.109	0.106
긴장형두통과 군발성두통의 차이	0.479	0.453	0.402	0.133	0.111	0.093	0.086	0.082	0.077	0.073
군발성두통과 편두통의 차이	0.159	0.155	0.150	0.137	0.132	0.126	0.119	0.118	0.114	0.113

4. 결론

본 연구에서는 구간값 퍼지 데이터를 이용한 의학진단에서 진단 척도로 흔히 활용되는 해밍 거리의 특성에 관해 탐색해 보았다. 환자가 갖는 증상의 수가 많을 경우에는 각 거리들간에 특별한 관계가 나타나지 않으나 증상의 수가 적은 경우에는 각 거리들간에 역상관 형태의 관계를 찾을 수 있었다. 또한 해밍 거리들의 차이값은 일반적인 통계 분포를 따르지 않는 형태이기 때문에 가상 환자에 대한 시뮬레이션을 통하여 그 차이의 유의성 검정을 위한 임계값을 계산하였다.

향후 해밍 거리의 특성을 세밀하게 파악하기 위한 추가적인 연구가 더 필요하다. 예를 들어, 그림 3.1에서 증상의 수가 9개일때 해밍 거리의 차이값들이 뚜렷한 군집을 형성하고 있다. 이 군집들과 증상 및 질병의 출현 형태 사이의 관계는 매우 의미있는 정보이다. 또 해밍 거리는 표 2.4와 표 2.5의 IIFWAA 값을 통해 구해지므로 해밍 거리와 이들 변수간의 관계 파악 또한 유용한 정보가 될 것으로 여겨진다.

참고문헌

Adlassnig, K. P. (1986). Fuzzy set theory in medical diagnosis. *IEEE Transactions on Systems, Man and Cybernetics*, **16**, 260-265.

Ahn, J. Y., Han, K. S., Oh, S. Y. and Lee, C. D. (2011). An application of interval-valued intuitionistic fuzzy sets for medical diagnosis of headache. *International Journal of Innovative Computing, Information and Control*, **7**, 2755-2762.

Atanassov, K. (1986). Intuitionistic fuzzy sets. *Fuzzy Sets and Systems*, **20**, 87-96.

De, S. K., Biswas, R. and Roy, A. R. (2001). An application of intuitionistic fuzzy sets in medical diagnosis. *Fuzzy Sets and Systems*, **117**, 209-213.

Guo, X., Zhang, H. and Chang, Z. (2010). Image thresholding algorithm based on image gradient and fuzzy set distance. *ICIC Express Letters*, **4**, 1059-1064.

Hamming, R. W. (1950). Error detecting and error correcting codes. *Bell System Technical Journal*, **26**, 147-160.

Innocent, P. R. and John, R. I. (2004). Computer aided fuzzy medical daignosis. *Information Sciences*, **162**, 81-104.

Jo, J. (2011). A statistical analysis of the fat mass experimental data using random coefficient model. *Journal of the Korean Data & Information Science Society*, **22**, 287-296.

Jo, J. and Baik, J. W. (2010). A statistical analysis on the selection of the optimal covariance matrix pattern for the cholesterol data. *Journal of the Korean Data & Information Science Society*, **21**, 1263-1270.

Joung, K. H. and Chung, S. S. (2011). Health-related quality of life among home-dwelling people with arthritis in Korea: Comparative study of osteoarthritis and rheumatoid arthritis. *Journal of the Korean Data & Information Science Society*, **22**, 555-563.

Kammerdiner, A., Krokhmal, P. A. and Pardalos, P. M. (2010). On the Hamming distance in combinatorial optimization problems on hypergraph matchings. *Optimization Letters*, **4**, 609-617.

Kim, Y. H., Kim, S. K., Oh, S. Y. and Ahn, J. Y. (2007). A fuzzy differential diagnosis of headache. *Journal of the Korean Data & Information Science Society*, **18**, 429-438.

Liang, Z. and Shi, P. (2003). Similarity measures on intuitionistic fuzzy sets. *Pattern Recognition Letters*, **24**, 2687-2693.

Park, C. and Kim, T. Y. (2010). Order selection method for clinical pathway development in acute appendectomy. *Journal of the Korean Data & Information Science Society*, **21**, 43-50.

Park, J. H., Lim, K. M., Park, J. S. and Kwun, Y. C. (2008). Distances between interval-valued intuitionistic fuzzy sets. *Journal of Physics: Conference Series*, **96**.

Sanchez, E. (1976). Resolution of composite fuzzy relation equations. *Information and Control*, **30**, 38-48.

Sanchez, E. (1979). Medical diagnosis and composite fuzzy relations. In *Advances in fuzzy set theory and applications*, edited by Gupta, M. M., Ragade, R. K. and Yager R. R., 437-444.

Szmidt, E. and Kacprzyk, J. (2001). Intuitionistic fuzzy sets in intelligent data analysis for medical diagnosis. *Lecture Notes in Computer Science*, **2074**, 263-271.

Wang, W and Xin, X. (2005). Distance measure between intuitionistic fuzzy sets. *Pattern Recognition Letters*, **26**, 2063-2069.

Zadeh, L. A. (1969). Biological applications of the theory of fuzzy sets and systems. *Proceedings of an International Symposium on Biocybernetics of the Central Nervous System*, 199-206.

Zeng, W., Yu, F., Yu, X., Chen, H. and Wu, S. (2009). Entropy of intuitionistic fuzzy set based on similarity measure. *International Journal of Innovative Computing, Information and Control*, **5**, 4737-4744.

Journal of the Korean Data & Information Science Society
2012, **23**(2), 227–234

한국데이터정보과학회지

On the characteristics of the Hamming distances in medical diagnosis[†]

Jeong Yong Ahn[1]

Department of Statistics, Chonbuk National University

Received 23 December 2011, revised 13 January 2012, accepted 25 January 2012

Abstract

Hamming distances in medical science are used for the diagnosis of diseases. The differences of the distances, however, are often very small, and is not in the general statistical form such as normal or chi-square distribution. In this study, we explore the characteristics and significance of the differences of Hamming distances generated in medical diagnosis.

Keywords: Hamming distance, interval-valued fuzzy data, medical diagnosis.

[†] This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2011-0008954).
[1] Professor, Department of Statistics (Institute of Applied Statistics), Chonbuk National University, Jeonbuk 561-756. E-mail: jyahn@jbnu.ac.kr