

의학 진단 및 생물 정보학 분야에서 퍼지 접근법의 적용에 관한 연구[†]

정혜영¹

¹서울대학교 기초교육원

접수 2018년 10월 11일, 수정 2018년 11월 18일, 게재 확정 2018년 11월 19일

요약

의료 진단 및 생물 정보학 분야에서 일어지는 자료에 존재하는 관계, 속성 및 개체들은 근본적으로 퍼지 (fuzzy)하다. 이러한 자료를 다루기에 적합한 이론이 퍼지 집합 이론이다. 퍼지 집합 이론은 경계의 불확실성 및 분류의 불확실성을 다룰 수 있도록 창안된 이론으로 명확한 경계와 분류를 가지는 기준의 집합 이론을 포함하는 확장개념으로 여겨진다. 본 논문에서는 퍼지 집합 이론에 기반 한 퍼지 접근법이 의학 진단 및 생물 정보학 분야의 자료에 어떻게 적용될 수 있는지 살펴보고 지금까지 성공적으로 적용되어 온 다양한 사례에 대해 소개하고자 한다.

주요 용어: 불확실성, 생물 정보학, 의학 진단, 퍼지 집합 이론.

1. 서론

우리의 현실 세계는 흑과 백, 0과 1, 참과 거짓으로 표현할 수 없는 부드러움 (softness)으로 표현할 수 있는 경우가 대부분이다. 두 가지 값을 사용하는 부울 논리의 딱딱함 (hardness)으로 이러한 부드러움의 현실 세계를 모델링 하게 되면 연속적인 공간에서 두 가지 값을 가지는 부울 논리로의 변환이 일어나게 되는데 이 과정에서 중요한 정보와 정밀도가 손실된다 (Barro와 Marin, 2002). 자료를 퍼지 집합으로 표현하는 것은 부정확 (imprecision)하거나 불확실 (uncertainty)한 것 또는 애매하고 모호한 것 (vagueness)을 부울 논리로 구분하지 않고 자연스럽게 그 정도에 대한 값을 매긴 것으로 이를 통해 확실성과 정확성을 구하는 작업이다. 한 원소가 특정 집합에 속하느냐 속하지 않느냐 하는 부울 논리로는 우리 삶의 극히 일부분만을 표현할 수 있을 뿐이다. 실제 우리의 생활 현상을 표현하기 위해서는 자연 언어로 이루어진 자료를 수학적으로 표현할 수 있는 확장된 집합의 개념이 필요하다. ‘두통이 심하고 열이 나며 오한이 드는 환자가 방문할 경우 몸을 시원하게 해 주되 따뜻함을 유지할 수 있도록 얇은 이불을 덮어주십시오’라는 진술을 처리하고 모델링하기 위해서는 고도의 프로그래밍 기술 및 부울 논리 이상의 무엇인가가 요구된다. 이러한 자연 언어를 처리하고 자료가 지닌 모호성을 표현하기 위해 Zadeh (1965)는 퍼지 집합의 개념을 도입한다. 퍼지 집합은 원소가 속하는 정도가 얼마인가를 다루는 집합으로 보통 우리가 다루는 집합의 일반화이고 보통 집합은 퍼지 집합의 특수한 경우가 된다.

[†] 이 성과는 2018년도 과학기술정보통신부의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No. 2017R1C1B1005069).

¹ (08826) 서울특별시 관악구 관악로1, 서울대학교 기초교육원, 강의부교수.

E-mail: hyjunglove@snu.ac.kr

이러한 퍼지 집합의 표현이 자연스럽게 느껴지는 가장 대표적인 분야가 바로 의학 진단 및 생물 정보학 분야이다. 우리는 질병의 메커니즘에 대한 완전한 이해가 부족하고 건강(또는 병) 상태에 관한 완전한 정보를 얻을 수 없으며 정상적인(그리고 비정상적인) 범위의 부정확성, 그리고 의학적 개념과 용어와 관련된 내재적 모호성 및 애매 모호성을 지니고 있다. 또한 질병 진단에도 명확한 경계를 내릴 수 없는 부정확성과 불확실성이 존재하게 되며 같은 증상인 경우에도 다른 질병을 나타내거나 그 반대인 경우도 많이 나타난다. 이처럼 의학에서 나오는 자료는 거의 대부분이 명확한 경계를 가지고 나눌 수 있는 자료가 아니며 반드시 하나의 집합에만 정확하게 분류되지 않는 자료이므로 거의 대부분의 자료가 퍼지 자료라고 해도 과언이 아닐 것이다. 생물 정보학에서도 실제 자료를 얻는 과정에서 경계의 명확성이 없는 경우가 많고 생물학적으로 그 역할들에 대해서 명확하지 않거나 여러가지 기능을 동시에 하는 경우가 많기에 이 또한 대부분 퍼지 자료인 것이다. 정밀의학을 향해서 나아가고 있는 지금의 현실에서 인간의 사고와 감정을 그대로 반영하는 퍼지 이론을 이용한 퍼지 접근법으로 자료를 분석하고 해석하는 일은 이분법적 구분에 기반한 기존의 접근법으로 해결할 수 없거나 혹은 잘 표현할 수 없는 부분을 해결해 줄 수 있는 바람직한 해법이 될 것이다.

2. 퍼지 집합 및 퍼지 논리

퍼지 집합은 수학자이자 전산학자인 캘리포니아 버클리 대학교 전산학과 교수 Zadeh (1965)에 의해 처음 제시된 이후로 많은 연구자에 의해 연구가 이어져 왔다 (Negoita와 Ralescu, 1975; Yager와 Filev, 1994; Klir와 Yuan, 1996; Reznik, 1997; Pedrycz와 Gomide, 2007; Nguyen와 Walker, 2000). 퍼지 집합은 한 원소가 하나의 집합에 속하는가? 속하지 않는가? 라는 이진 논리가 아니라 원소가 그 집합에 속하는 정도가 얼마인가를 다루기 위한 집합이다. 즉 속하면 0, 속하지 않으면 1로 소속의 정도를 표현하는 소속함수의 값을 속하는 정도에 따라서 [0, 1] 사이의 다양한 값을 매길 수 있도록 보통 집합을 확장한 집합의 개념으로 퍼지 집합은 보통 집합의 일반화이고 보통 집합은 퍼지 집합의 특수한 경우에 해당한다.

‘열’에 대해 생각해보자. 보통 집합이 열이 38도 이상일 경우 열이 나는 것으로 분류를 하게 된다고 할 때, 37.9도는 36.5도와 동일하게 열이 나지 않는 것으로 분류된다. 이것을 퍼지 집합으로 표현하게 되면 38도 이상부터는 정확히 열이 나는 것으로 소속정도 1의 값을 가지지만 37.9인 열도 36.5도 보다는 ‘열이 난다’는 집합에 대해 높은 소속 정도를 가지는 것으로 표현할 수가 있게 된다. 우리는 ‘열’을 Figure 2.1과 같이 ‘absent’, ‘medium’, ‘high’의 세 가지 언어적 용어 (linguistic term)를 가지는 언어적 변수 (linguistic variable)로 표현할 수 있다. 37도를 넘지 않을 때는 열이 없는 것으로 간주한다면 37도 보다 낮은 열은 정도의 차이가 없이 열이 없음을 의미하는 ‘absent’라는 집합에 소속 될 가능성을 모두 1로 표현할 수 있다. 하지만 37도와 39도 사이의 열이 날 경우 고열은 아니지만 어느 정도 열이 난다는 것을 의미하게 되나 37.1도의 열과 38도의 열이 정도의 차이가 없이 동일하게 ‘medium’이라고 표현되는 것은 자연스럽지 않다. 따라서 우리는 ‘medium’이라는 퍼지 집합을 정의하여 37.1도의 열과 38도의 열에 대한 소속 정도를 다르게 표현할 수 있게 된다. 뿐만 아니라, 38도의 열은 1의 소속정도를 가지고 ‘medium’의 집합에 속하지만 아직 고열은 아니므로 ‘high’의 집합에는 소속 정도 0를 가진다. 하지만 38도가 넘는 열이 나게 되면 점점 ‘medium’집합에 소속될 정도는 작아지고 ‘high’집합에 소속될 정도는 커지게 된다. 결국 열의 정도에 따라 구분을 해줄 뿐 아니라 한 사람이 여러 집합에 동시에 소속되는 것을 허용함으로써 더욱 현실적으로 실제 현상을 반영할 수 있게 되는데 이것이 퍼지 집합 이론을 사용해야 하는 근본적인 이유이다. 이처럼 실수로 주어지는 자료를 언어적 표현으로 나타내어 퍼지 집합에 의해 표현하는 것은 퍼지 집합의 응용에 있어 중요한 역할을 하며 특히 approximate reasoning에 있어서는 필수적이다. ‘젊다’, ‘따뜻하다’, ‘많이’, ‘적당히’ 등과 같은 언어적인 표현도 퍼지 집합으로 표현하는

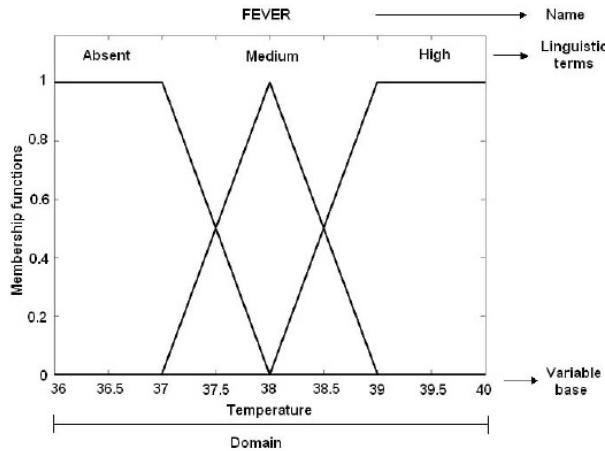


Figure 2.1 An example of a linguistic variable (Massad *et al.*, 2009)

것이 가능하기에 우리 삶의 다양한 생각과 의견을 자연스럽게 분석할 수 있는 도구로 퍼지 집합은 사용되어지고 있다.

Figure 2.2와 2.3은 ‘건강하다’와 ‘건강하지 않다’라는 분류 과정을 보통 집합과 퍼지 집합으로 나타낸 그림이다. 제시 된 그림을 통해 ‘건강하다’와 ‘건강하지 않다’는 명확한 경계를 가지고 나눌 수 있는 개념이라기 보다는 서로 겹치는 부분이 있으며 건강한 정도 (또는 건강하지 않은 정도)를 나타내는 퍼지 집합으로 표현하는 것이 더욱 자연스러운 표현임을 알 수 있게 된다. 특히 의학과 생물 정보학 분야의 경우, 사용되어지는 각 개념과 개념간의 관계에 대한 명확한 정의나 설명을 제공하는 것이 불가능한 경우가 많기 때문에 보통 집합보다 퍼지 집합을 사용하는 것이 적합한 경우가 많다.

Figure 2.4는 상파울로 병원의 역학자들이 호흡 곤란을 묘사하기 위해 사용한 포스터이다. 이 그림을 통해 호흡 곤란의 정도는 ‘absent’, ‘slight’, ‘moderate’ and ‘severe’와 같은 언어적 용어를 갖는 언어적 변수로 표현되는 것이 자연스러운 표현임을 알 수 있다. 이를 각각은 0, 1, 2, 3의 숫자로 단계적으로 변하는 값을 나타내며 막대기의 색깔이 각각의 상황에 따라 연속적으로 점진적으로 변화됨을 나타내고 있다. 이처럼 ‘호흡 곤란’이라는 의학적 용어의 값이 가지는 점진적인 전환은 용어 자체의 모호함을 나타내며 퍼지 집합으로 표현될 수 있음을 알 수 있다.

Sadegh-Zadeh (2015)는 의학 지식은 피할 수 없는 불확실성에 의해 특징 지워진다고 주장한다. 이 불가피한 불확실성에 대한 많은 이유가 존재하지만 가장 중요한 이유는 정보 부족, 부정확한 정보 및 서로 상반된 결과를 얻을 수 있는 모순된 특징으로부터 기인하는 필연적인 모호함에 있다. 퍼지 논리는 이러한 모호함을 해결하며 단독으로 혹은 하이브리드 방식으로 응용되고 있다. Figure 2.5는 퍼지 논리를 사용한 퍼지 추론 시스템의 과정을 보여준다. 퍼지 추론 시스템은 시스템의 모호성을 처리하기 위해 퍼지 집합을 사용하는 규칙 기반 시스템 (Klir 등, 1988)으로 실수로 입력되는 값을 퍼지값으로 변환한 후 퍼지 추론 과정을 통해 결과값으로 퍼지값을 얻고 다시 실수로 그 값을 변환해주는 일련의 과정을 의미한다.

실수로 얻은 자료를 퍼지화하여 퍼지 접근법을 쓰거나 처음부터 퍼지 집합으로 표현되어야 하는 용어에 이르기까지 의학 및 생물 정보학 분야에서 사용하는 개념들은 퍼지 집합으로 표현하는 것이 더 적합

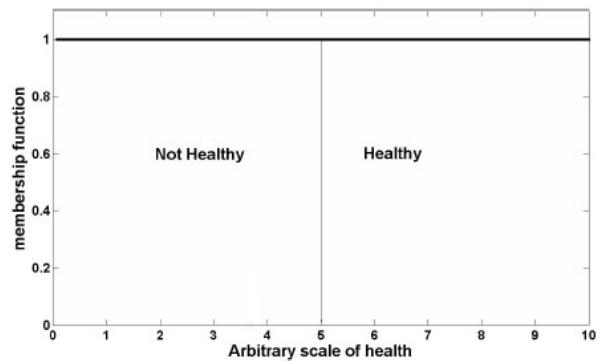


Figure 2.2 Sets to illustrate the health classificatory process in the classical approach (Massad *et al.*, 2009)

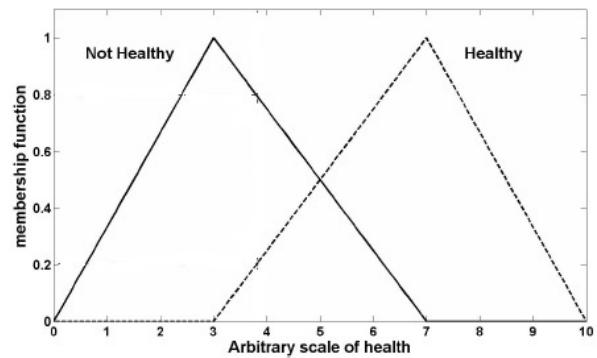


Figure 2.3 Sets to illustrate the health classificatory process in the fuzzy approach (Massad *et al.*, 2009)

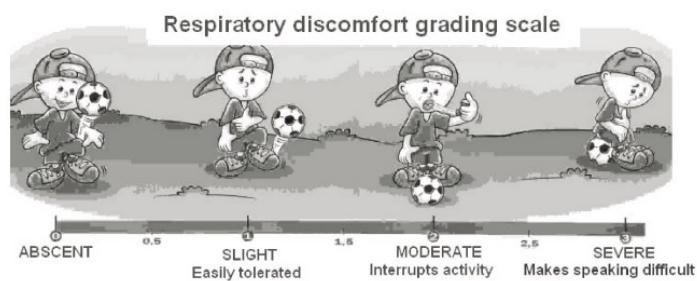


Figure 2.4 Poster in use by epidemiologists in the Sao Paulo Hospital of Clinics to describe the breathing discomfort (Massad *et al.*, 2009)

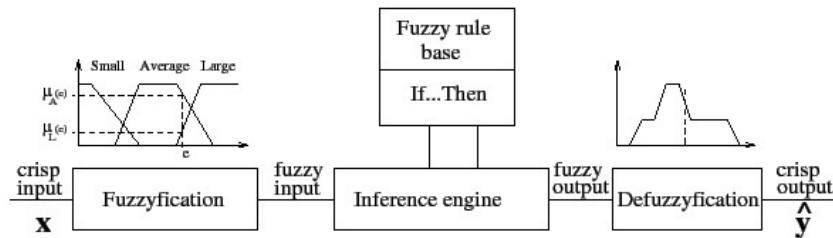


Figure 2.5 The process of fuzzy inference system
(<http://www7.inra.fr/mia/M/fispro/fisprodocen/QUICKSTART/img31.png>)

하고 자연스러우며 정보의 손실이 덜하다는 것을 앞서 제시된 예시들을 통해 알 수가 있었다. 이제 퍼지 집합 및 퍼지 논리를 이용한 의학 진단 및 생물 정보학에서의 퍼지 접근법들을 구체적으로 살펴보기로 하자.

3. 의학 진단 및 생물 정보학 분야에서 퍼지 접근법의 적용

의학 진단을 위해 가장 많이 사용되는 분야인 데이터 마이닝 (Hong 등, 2018)은 불완전한 자료로부터 다양한 유형의 수치 또는 기호를 얻게 되거나, 부정확하고 모호하면서 이질적인 데이터를 함께 다루게 되는 경우가 많이 발생한다. 퍼지 논리는 이렇게 주어지는 잡다한 형태의 자료를 표현할 수 있기 때문에 의학 진단 분야에 유용하다. 퍼지 집합과 퍼지 논리를 사용한 데이터 마이닝 기반의 퍼지 접근 방법으로는 퍼지 서포트 벡터 머신, 퍼지 클러스터링, 퍼지 신경망 분석등이 있으며 기존의 거의 모든 데이터 마이닝 기술이 퍼지 버전으로 확장되어 경계의 모호함을 가지거나 언어적 표현의 자료들, 하나의 원소가 다양한 역할을 가질 때 등의 기존 방법이 해결하기 쉽지 않는 영역에서 그 역할을 감당하고 있다. 이 절에서는 퍼지 접근법이 의학 진단을 위해 사용된 구체적인 사례를 제시해보고자 한다.

3.1. 의학 진단 분야에서 퍼지 접근법 기반의 연구들

의학 진단 분야에서의 불확실성의 근원은 다음과 같이 분류될 수 있다 (Abbad 등, 2001).

- 환자 및 가족이 제공한 환자에 대한 부정확한 과거 병력
- 의사가 검사를 통해 환자로 부터 얻는 자료의 많은 경우에 존재하는 정상인과 그렇지 않은 환자 사이의 불명확한 경계
- 환자의 병에 대한 과장된 혹은 절제된 증상
- 환자는 실제 수학적으로 명확히 정의되는 방식으로 상황을 표현하는 것이 아니라 부정확하고 모호한 용어와 언어를 사용
- 진단을 위해 미리 설계된 범주화된 자료 형태의 시스템을 사용하게 될 때 분류의 어려움
- 경계가 모호한 의료 영상 자료들

의학 진단 분야에서 가장 퍼지 접근법이 활발하게 적용되는 분야가 의료 영상 자료의 분석이다. 의료 영상은 의학적 진단을 지원하는 강력한 지원 요소 중 하나인데 영상에는 경계가 모호하거나 노이즈로 인해 분류가 쉽지 않은 경우가 많다. 이를 위해 퍼지 분류, 퍼지 클러스터링, 퍼지 규칙 기반 방법 및 퍼지

패턴 매칭 방법등이 사용되고 있다. 퍼지를 이용한 영상 분석이 기존의 분석 방법보다 경계를 더욱 명확하게 구분해주는 성공적인 사례들이 보고 되고 있으며 뇌종양의 진단 지원, 방사선 영상의 분류, 심전도 분석에 의한 부정맥의 분류 등에서 퍼지 접근법이 사용되었다 (Comas 등, 2011; Begum 등, 2011; Kobashi 등, 2002; Kerre 등, 2013; Ganasala 등, 2015).

질병의 진단 및 예후에 많이 사용되는 데이터 마이닝 기술에 대한 Kolce와 Frasher와 (2012)의 문헌 검토에서 예후의 예측에는 베이지안 알고리즘 및 퍼지 알고리즘이 가장 우수한 것으로 기술되어 있다. Hedeshi 등 (2011)은 Particle swarm optimization (PSO) 및 고급 데이터 마이닝 기술을 의학에서 퍼지 논리와 결합하여 암, 특히 뇌종양과 유방암을 탐지하기 위한 목적으로 사용하였다. Yardimci (2009)에 의하면 신경-퍼지 (neuro-fuzzy)의 하이브리드 방법이 가장 많이 사용되는 데이터 마이닝 기법이며 임상과학 58%, 기초과학 25%, 진단과학 17%로 응용 사례의 비율이 보고 되고 있다.

Klir과 Yuan (1996)은 퍼지 로직에 기반한 의학적 진단 시스템인 CADIAG 시스템을 모델링 하였고 좀 더 발전된 형태의 진단 시스템 CADIAG-2가 Adlassnig과 KolarZ (1982), Adlassnig 등 (1984, 1985), 그리고 Adlassnig (1986)에 의해 소개 되어졌다. Leite 등 (2000)은 이 진단 시스템의 향상된 버전인 CADIAG-II/RHEUMA에 의해 류마티즘을 진단하고 환자의 증상을 다른 특정 진단과 관련시키는 과정을 소개하였다. 이처럼 퍼지 진단 시스템의 실제 활용에 이르기까지 퍼지 접근법은 의학적 진단 분야에서 활발히 적용되어지고 있다.

3.2. 생물 정보학 분야에서 퍼지 접근법 기반의 연구들

지금까지 거의 모든 생물 정보학 분야의 문제는 집합 이론에 근거한 확률적 불확실성을 다루는 통계적 방법으로 구현되어져 왔다. 그러나 실제로 얻어지는 자료들은 확률적 불확실성 뿐만 아니라 다음과 같은 현상들로 인해 기존의 확률적 접근법과 퍼지 접근법이 함께 고려되어야 한다 (Xu 등, 2006).

- 생물 시스템에서의 본질적인 모호성
- 생물학적 객체의 다중 역할
- 생물학적 현상에 대한 fuzzy descriptions

생물학적 시스템에서 많은 프로세스가 결정론적이기 보다는 본질적으로 모호하다는 증거가 점점 더 많이 발견되고 있다. 예를 들어 내성의 결과인 면역계는 자기 자신의 항원을 모두 퍼지 방식으로 인식할 수 있는 것으로 밝혀졌다. 이러한 면역 시스템의 퍼지 특징은 암과 같은 자가 면역 질환의 기전을 밝혀 줄 수 있다. 또한 확률 이론과 퍼지 이론을 결합하여 세포 행동의 가변성을 설명할 수 있는 새로운 유형의 제어 메커니즘을 제공할 수도 있다 (Samoilov 등, 2005). 생물학적 개체는 여러가지 역할을 수행할 수 있으므로 퍼지 소속함수값이 생기게 된다. 예를 들어 베타 카테닌은 다기능 단백질로서 세포 간 부착과 세포내 신호 전달에 중요한 역할을 한다 (Steinberg 등, 1999). 실제로 생물학적 자료를 사용하여 유전자를 클러스터링 할 때, 유전자에 대한 퍼지 소속함수 또는 퍼지 클러스터링이 더욱 높은 설명력을 보여주기도 한다. 이 경우 하나의 유전자가 두 개 이상의 클러스터에 동시에 존재할 수 있으며 각 클러스터에 부분적으로 혹은 전체적으로 속할 수 있는 소속함수값이 존재하게 된다 (Sasik 등 2001).

생물 정보학에서 유전자 특성은 모호하고 이를 분류하는 유전자 온톨로지 및 유전자와 유전자 온톨로지 사이의 매핑 또한 모호한 관계를 가질 수 있다. Dweiri 등 (2006)에서는 관계형 퍼지 c-means (NERFCM)을 사용하여 집단 내의 유전자 산물에 퍼지 소속함수값을 할당하였고 Roychowdhury 등 (2004)에서는 유전자 온톨로지 유사성에 의해 마이크로어레이 실험에서 기능적으로 관련이 있는 발현 유전자 클러스터를 발견하였다. 온톨로지 기반에 대한 보다 일반적인 퍼지 접근법은 Torres 등 (2006)에 제시되어 있다. 수많은 연구에도 불구하고 단백질의 2차 구조를 정확하게 예측하는 것은 여전

히 어려운 문제로 남아있다. 퍼지 클러스터링 방법을 사용하여 단백질 2차 구조 예측과 단백질 서열로부터 각 아미노산의 용매 접근성을 예측하는 방법이 Sim 등 (2005)에 의해 제시되었다.

Jung 등 (2016, 2017)은 유전자간 상호작용을 밝혀내기 위해 사용되는 MDR (multifactor dimensional reduction) 방법에 퍼지 집합을 접목하여 Fuzzy-MDR 방법을 제시하였다. 기존의 MDR 방법은 특정한 기준을 근거로 위험군과 위험하지 않은 군으로 각 유전자 조합을 분류하지만 이 방법은 특정한 기준값 근처에 있는 값을 구분하지 못하는 단점이 있다. 이를 해결하기에 적합한 이론이 퍼지 이론이다. Fuzzy-MDR 방법은 특정한 기준을 중심으로 떨어진 정도에 따라 다른 소속정도를 가지고 위험군과 위험하지 않은 군에 동시에 소속될 수 있도록 허용함으로써 더 많은 정보를 반영한 판정을 하게 해준다.

유전자의 발현 정도도 사실은 명확한 경계를 갖지 않는 경우가 많기 때문에 경계의 흐릿함을 반영하는 퍼지 접근법을 사용하여 유전자의 발현 정도에 따른 차이를 분석하는 방법들이 더 적합할 수 있다. Hosseini 등 (2018)은 생물학적 과정에서 유전자의 기능적 관계를 밝혀내기 위해 유전자 발현 정도에 따른 클러스터링 방법으로 퍼지 클러스터링을 사용하였다. 이처럼 생물 정보학 분야의 부정확성, 불확실성, 애매함을 표현하기 위한 퍼지 접근법의 잠재력을 앞서 제시한 예제 뿐만 아니라 더욱 다양한 실제 상황을 구현하는데 있어서 여전히 진행형이다.

4. 실제 응용 사례 예시 및 유용한 통계 패키지

4.1. 실제 응용 사례

인슐린 저항성 지수에 영향을 주는 유전자 간 교호작용을 찾기 위해, SNP (single nucleotide polymorphism, 단일 염기 다형성)의 조합 중 가장 환자군과 대조군을 잘 구분하는 조합을 찾는 비모수적인 통계적 방법인 MDR 방법이 많이 사용되어져 왔다. 먼저, MDR을 사용하고자 할 때, 우리는 환자군 (case)과 대조군 (control)의 비를 이용하게 되는데 전체 집합에 대한 비와 각 SNP 조합에 대한 비를 비교해서 각 SNP의 조합을 높은 위험군과 낮은 위험군으로 분류하게 된다. Figure 4.1을 보자.

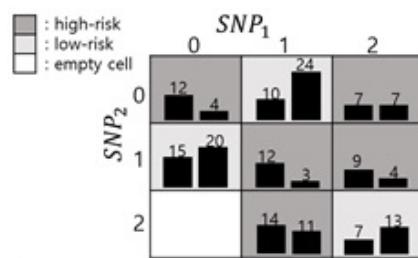


Figure 4.1 Measure membership degrees of multifactor classes using MDR

n_{i0} 를 SNP1과 SNP2의 조합에서 나타나는 i 번째 셀에 존재하는 대조군의 수라 하고 n_{i1} 를 SNP1과 SNP2의 조합에서 나타나는 i 번째 셀에 존재하는 환자군의 수라 할 때, 전체 환자군의 수/전체 대조군의 수를 1이라고 한다면 1보다 큰 값을 가지는 셀은 질병에 대한 높은 위험군으로 분류하게 된다. 짙은 회색은 높은 위험군을 뜻하고, 연한 회색은 낮은 위험군을 뜻한다. 이 때, 첫번째 셀에 대한 비 12/4와 8번째 셀에 대한 비 14/11는 차이가 큰 값이지만 이들의 차이가 반영되지 못하고 그저 높은 위험군에 속한다고 분류하게 된다. 우리는 여기에서 퍼지 집합의 적용 가능성을 생각할 수 있다. 각 셀은 정확하게 높은 위험군, 낮은 위험군의 두 그룹으로 분류하기 힘들 뿐 아니라 환자군/대조군의 비도 각 그룹에 대해 어느 정도의 강도로 속하게 되는지를 나타내는 척도가 되므로 그 정보도 충분히 반영이 되어야 한

다. 이를 위해 높은 위험군과 낮은 위험군이라는 두 그룹을 퍼지 집합으로 표현하여 각 셀이 이 두 그룹에 동시에 속될 수 있도록 허용하고, 각 셀에서 얻어지는 환자군과 대조군의 비의 크기를 반영하고자 한다. 이제 우리는 12/4의 비를 갖는 셀은 높은 위험군에 속하는 정도가 14/11보다 높고, 낮은 위험군에 속하는 정도는 14/11보다 높지 않도록 퍼지 집합에 속하는 정도를 매길 수 있게 된다. Figure 4.2은 높은 위험군에 속하는 정도를 중심으로 그린 그림이다. 높은 위험군과 낮은 위험군에 속하는 정도의 합이 1이므로 1에서 높은 위험군에 속하는 소속 정도를 빼면 낮은 위험군에 속하는 정도가 된다. Figure 4.2에서 하얀색은 환자군과 대조군의 수가 같은 경우를 뜻한다. MDR에서는 이러한 셀은 높은 위험군으로 일반적으로 분류를 했지만 Fuzzy-MDR에서는 각 위험군에 속하는 정도가 0.5씩 주어지도록 함으로써 우리의 자연스런 사고를 반영하게 된다는 것을 알 수 있다. 이 과정을 통해 판정된 위험군에 대한 결과를 바탕으로 분류 정확도를 비교하여 인슐린 저항성 지수에 영향을 주는 SNP의 조합을 찾게 된다 (Jung, 2016, 2017).

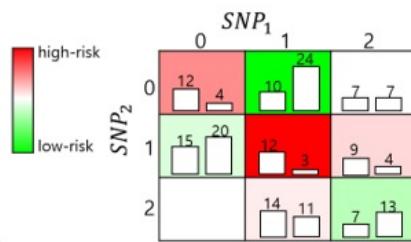


Figure 4.2 Measure membership degrees of multifactor classes using Fuzzy-MDR

다음으로는 퍼지 논리를 사용한 방법들에 대한 이해를 돋고자 퍼지 논리 규칙의 예를 하나 들고자 한다.

- Rule 1: 만약 암의 크기가 작다면 암에 대한 위험도는 낮다.
- Rule 2: 만약 암의 크기가 적당하다면 암에 대한 위험도는 낮다.
- Rule 3: 만약 암의 크기가 적당하다면 암에 대한 위험도는 높다.

여기서 ‘작다’, ‘적당하다’, ‘낮다’, 그리고 ‘높다’와 같은 언어적인 표현을 가능케 하기 위해 퍼지 집합이 필요하게 되고 이렇게 퍼지 집합으로 논리를 표현하는 것을 퍼지 논리라고 부르게 된다. 의학 진단 및 생물 정보학 분야의 많은 자료들은 실제로는 연속된 수치를 얻는데 어떠한 정해진 범주로 측정을 하기 때문에 이 범주 기준을 기반으로 분석을 하게 될 경우 자료의 손실이 발생할 수 있다. 실제로 얻어진 연속된 수치를 범주화 된 자료로 바꾸어 분석에 사용하는 것이 아니라 자료를 퍼지 집합으로 표현하고 이렇게 퍼지 논리에 의해 그 관계를 표현하는 것이 더 자연스러울 수 있다.

이렇게 퍼지 집합이론을 적용한 통계적 방법론을 사용할 때, 우리의 사고 과정을 더 자연스럽게 반영하게 되고 0 또는 1의 값으로 소속 정도를 매겨서 속하지 않는다/속한다라고 분류하는 과정에서 발생하는 정보의 손실을 줄일 수 있게 된다.

4.2. 유용한 통계 패키지

클러스터링을 비롯한 머신 러닝 기법 등과 같은 통계적 분석 방법의 대부분은 0,1을 분류 기준으로 하는 집합 이론에 근거하고 있다. 이 0,1의 분류 기준을 구간 [0, 1]로 확장하고, 하나의 원소가 여러 그룹

에 동시에 소속될 수 있도록 허용하는 퍼지 집합 이론을 접목하여 기존 방법의 확장 된 개념으로 퍼지 분석 방법을 사용하게 된다. 퍼지 분석 방법들이 실제 사례에 응용될 수 있도록 유용한 R패키지들을 소개하고자 한다.

- sets : 퍼지 논리를 사용해서 퍼지 규칙 기반 추론 시스템을 구현 할 수 있다.
- e1071: 퍼지 c-평균 알고리즘을 구현 할 수 있다.
- fugeR: 유전자 알고리즘을 사용해서 퍼지 논리 규칙으로 이루어진 예측 모형을 구현 할 수 있다.
- frbs: 클러스터링, 신경망, 공간 분할 등과 같은 방법을 통해 데이터를 학습시켜서 퍼지 규칙 기반 시스템을 구현 할 수 있다.
- fclust: 퍼지 클러스터링의 다양한 알고리즘을 구현 할 수 있다.
- ade4: 퍼지 대응 분석 및 퍼지 주성분 분석을 구현 할 수 있다.

5. 결론

퍼지 접근법은 근본적인 생물학적 매커니즘이 아닌 모호함을 표현하고 분석 및 예측을 하기 위한 방법으로 의학 및 생물 정보학에 존재하는 많은 문제를 해결하기에 적합하다. 물론 모든 경우에서 퍼지 접근법의 성능이 우수하지는 않을 수 있다. 우리는 퍼지 접근법을 사용하는 것이 더 효과적인 예시들을 이 논문에서 제시하였고 그 잠재력을 감안할 때 앞으로 더 많은 연구가 진행될 것으로 생각된다. 전문가의 의견과 의학 및 생물학적 정보의 결합도 정밀 의학을 구현하려는 21세기의 의료 시스템에서는 필수적으로 요구되고 있는데 출처가 다른 이러한 자료의 통합에도 퍼지 논리를 이용한 접근법은 좋은 해법이 될 수 있다. 퍼지 접근법의 의료 및 생물 정보학 분야에서의 응용 사례는 매년 증가하고 있다. 그 중 가장 많이 사용되는 접근법이 퍼지 논리와 인공 신경망의 혼합 모형이다 (Lee, 2015; Yardimci, 2009).

퍼지 접근법의 주요 응용 분야는 클러스터링 및 분류 연구, 패턴 인식 및 기능 선택에 대한 연구, 이미지 처리 연구 등이 있는 것으로 조사되었다. 즉 정밀 의학을 향한 모든 의사결정 과정 속에 존재하는 모호함을 처리하기 위해 퍼지 접근법이 사용되고 있음을 알 수 있다. 퍼지 접근법을 사용하게 됨으로 우리는 더 많은 정보를 분석에 사용할 수 있게 된다. 그러나 퍼지 접근법은 소속 함수의 주관적인 선택에 대한 문제, 퍼지 의사결정 시스템에 의해 주어지는 퍼지 소속 함수값에 대한 해석의 문제, 자료가 지닌 수 많은 차원을 해결하기 위한 컴퓨팅 기술의 문제 등이 여전히 해결해야 할 주제로 남아있다. 그럼에도 불구하고 퍼지 접근법은 아직도 많은 잠재력을 가지고 있기에 더욱 많은 의학 진단 및 생물 정보학 분야에 적용되기를 바라며 또한 정밀 의학의 구현을 향한 의료 시스템의 구축에 기존의 하드 컴퓨팅 방법과 퍼지 논리를 기반으로 한 소프트 컴퓨팅 방법의 혼합 사용에 대한 잠재적인 수요가 더욱 가시화되기를 기대한다.

References

- Abbad, M., von Keyserlingk, D., Linkens, D. and Mahfouf, M. (2001). Survey of utilisation of fuzzy technology in medicine and healthcare. *Fuzzy Sets Systems*, **120**, 331-49.
- Adlassnig, K. P., Scheithauer, W. and Grabner, G. (1984). Computer-assisted diagnosis and its application in pancreatic diseases. *Acta. Med. Austriaca*, **11**, 125-134.
- Adlassnig, K. P., Kolarz, G., Scheithauer, W., Effenberger, H. and Grabner, G. (1985). CADIAG: Approaches to computer-assisted medical diagnosis. *Computers in Biology and Medicine*, **15**, 315-335.
- Adlassnig, K. P. (1986). A survey on medical diagnosis and fuzzy subsets. *Approximate Reasoning in Decision Analysis*, 203-217.

- Adlassnig, K. P. and Kolarz, G. (1982). CADIG-2: Computer-assisted medical diagnosis using fuzzy subsets. *Approximate Reasoning in Decision Analysis*, 219-247.
- Barro, S. and Marin, R. (2002). *Fuzzy logic in medicine*, Physica.
- Begum, S. A. and Devi, O. M. (2011). Fuzzy algorithms for pattern recognition in medical diagnosis. *Assam University Journal of Science and Technology*, **7**, 1-12.
- Bellamy, J. E. (1997). Medical diagnosis, diagnostic spaces, and fuzzy systems. *Journal of the American Veterinary Medical Association*, **210**, 390-396.
- Comas, D. S., Meschino, G. J., Pastore, J. I. and Ballarin, V. L. (2011). A survey of medical images and signal processing problems solved successfully by the application of type-2 Fuzzy Logic. In *Journal of Physics: Conference Series*, **332**.
- Dweiri, F. T. and Kablan, M. M. (2006). Using fuzzy decision making for the evaluation of the project management internal efficiency. *Decision Support Systems*, **42**, 712-726.
- Ganasala, P. and Kumar, V. (2015). Multi-scale decomposition transform based approach for CT and MR image fusion. *Journal of Medical Imaging and Health Informatics*, **5**, 715-723.
- Hedeshi, N. G. and Abadeh, M. S. (2011). An expert system working upon an ensemble PSO-based approach for diagnosis of coronary artery disease. In *Biomedical Engineering (ICBME), 2011 18th Iranian Conference*, 249-254.
- Hong, S., Kang, D. and Choi, J. (2018). Analysis of domestic diabetes prevalence data using Bayesian spatially-dependent clustering models in regression coefficients. *The Korean Data & Information Science Society*, **29**, 633-644.
- Hosseini, B. and Kiani, K. (2018). FWCMR: A scalable and robust fuzzy weighted clustering based on MapReduce with application to microarray gene expression. *Expert Systems with Applications*, **91**, 198-210.
- Jung, H. Y., Leem, S., Lee, S. and Park, T. (2016). A novel fuzzy set based multifactor dimensionality reduction method for detecting gene-gene interaction. *Computational Biology and Chemistry*, **65**, 193-202.
- Jung, H. Y., Leem, S. and Park, T. (2018). Fuzzy set-based generalized multifactor dimensionality reduction analysis of gene-gene interactions. *BMC Medical Genomics*, **11**, 32.
- Kerre, E. E. and Nachtegael, M. (2013). Fuzzy techniques in image processing. *Physica*, **52**.
- Klir, G. J. and Folger, T. A. (1988). *Fuzzy sets, uncertainty, and information*, Prentice Hal.
- Klir, G. J. and Yuan, B. (1996). Fuzzy sets and fuzzy logic: Theory and applications. *Possibility Theory versus Probab. Theory*, **32**.
- Kobashi, S., Hata, Y. and Hall, L. O. (2002). Fuzzy information granulation of medical images. Blood vessel extraction from 3-D MRA images. *Physica*, 18-35.
- Kolce, E. and Frasher, N. (2012). *A literature review of data mining techniques used in healthcare databases*, ICT innovations.
- Lee, K. J., Lee, H. J. and Oh, K. J. (2015). Using fuzzy-neural network to predict hedge fund survival. *Journal of the Korean Data & Information Science Society*, **26**, 1189-1198.
- Leite, M. B. F., Bassanezi, R. C. and Yang, H. M. (2000). The basic reproduction ratio for a model of directly transmitted infections considering the virus charge and the immunological response. *Mathematical Medicine and Biology: A Journal of the IMA*, **17**, 15-31.
- Massad, E., Ortega, N. R. S., de Barros, L. C. and Struchiner, C. J. (2009). *Fuzzy logic in action: Applications in epidemiology and beyond*, Springer Science & Business Media.
- Negoita, C. V. and Ralescu, D. A. (1975). *Application of fuzzy sets to system analysis*, Birkhäuser Verlag, Basel und Stuttgart, Germany.
- Nguyen, H. T. and Walker, E. A. (2000). *A first course in fuzzy logic*, Chapman & Hall, USA.
- Pedrycz, W. and Gomide, F. (2007). *Fuzzy systems engineering: Toward human-centric computation*, John Wiley & Sons Publications, USA.
- Priya, D. K., Krithiga, S. R., Pavithra, P. and Kumar, J. R. (2015). Detection of leukemia in blood microscopic images using fuzzy logic. *Int J Eng Res Sci Technol*, **240**, 197-205.
- Reznik, L. (1997). *Fuzzy controllers*, Newnes, Great Britain.
- Roychowdhury, A., Pratihar, D. K., Bose, N., Sankaranarayanan, K. P. and Sudhahar, N. (2004). Diagnosis of the diseases using a GA-fuzzy approach. *Information Sciences*, **162**, 105-120.
- Sadegh-Zadeh, K. (2015). *Fuzzy logic*, In: Handbook of Analytic Philosophy of Medicine, Springer.
- Samoilov, M., Plyasunov, S. and Arkin, A. P. (2005). *Stochastic amplification and signaling in enzymatic futile cycles through noise induced bistability with oscillations*, Proc Natl Acad Sci, USA.
- Sasik, R., Hwa, T., Iranfar, N. and Loomis, W. F. (2001). Percolation clustering: A novel approach to the clustering of gene expression patterns in Dictyostelium development. *Pac Symp Biocomput.*, 335-47.

- Sharma, S. and Wasson, E. V. (2015). Retinal blood vessel segmentation using fuzzy logic. *Journal of Network Communications and Emerging Technologies*, **4**.
- Sim, J., Kim, S. Y. and Lee, J. (2005). Prediction of protein solvent accessibility using fuzzy k-nearest neighbor method. *Bioinformatics*, **21**, 2844-9.
- Steinberg, M. S. and McNutt, P. M. (1999). Cadherins and their connections: Adhesion junctions have broader functions. *Curr Opin Cell Biol*, **11**, 554-60.
- Torres, A. and Nieto, J. J. (2006). Fuzzy logic in medicine and bioinformatics. *BioMed Research International*.
- Wu, Y., Duan, H. and Du, S. (2015). Multiple fuzzy c-means clustering algorithm in medical diagnosis. *Technology and Health Care*, **23**, 519-527.
- Xu, D., Bondugula, R., Popescu, M. and Keller, J. (2006). Bioinformatics and fuzzy logic. In *Fuzzy Systems, IEEE International Conference*, 817-824.
- Yager, R. R. and Filev, D. P. (1994). *Essentials of fuzzy modeling and control*, Wiley-Interscience, USA.
- Yardimci, A. (2009). Soft computing in medicine. *Appl Soft Comput.*, **9**, 1029-43.
- Zadeh, L. A. (1965). Fuzzy sets. *Information and Control*, **8**, 338-353.
- Zadeh, L. A. (1969). *Biological applications of the theory of fuzzy sets and systems*, In: Proctor, L. D. (ed.) *Biocybernetics of the Central Nervous System*, Little Brown & Co., Boston.

On the applications of fuzzy approaches in medical diagnosis and bioinformatics[†]

Hye-Young Jung¹

¹Faculty of Liberal Education, Seoul National University

Received 11 October 2018, revised 18 November 2018, accepted 19 November 2018

Abstract

The relationships, properties, and objects in the data generated from medical diagnosis and bioinformatics are fundamentally fuzzy. Fuzzy set theory is an ideal framework to deal with such data. Fuzzy set theory is considered to be an extended set theory to deal with uncertainty of boundary and classification. In this paper, we illustrate how fuzzy approaches based on fuzzy set theory can be applied to data in medical diagnostics and bioinformatics with various examples.

Keywords: Bioinformatics, fuzzy set theory, medical diagnosis, uncertainty.

[†] This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MIST)(No. 2017R1C1B100506).

¹ Corresponding author: Associate teaching professor, Faculty of Liberal Education, Seoul National University, Seoul, 08826, Korea. E-mail: hyjunglove@snu.ac.kr