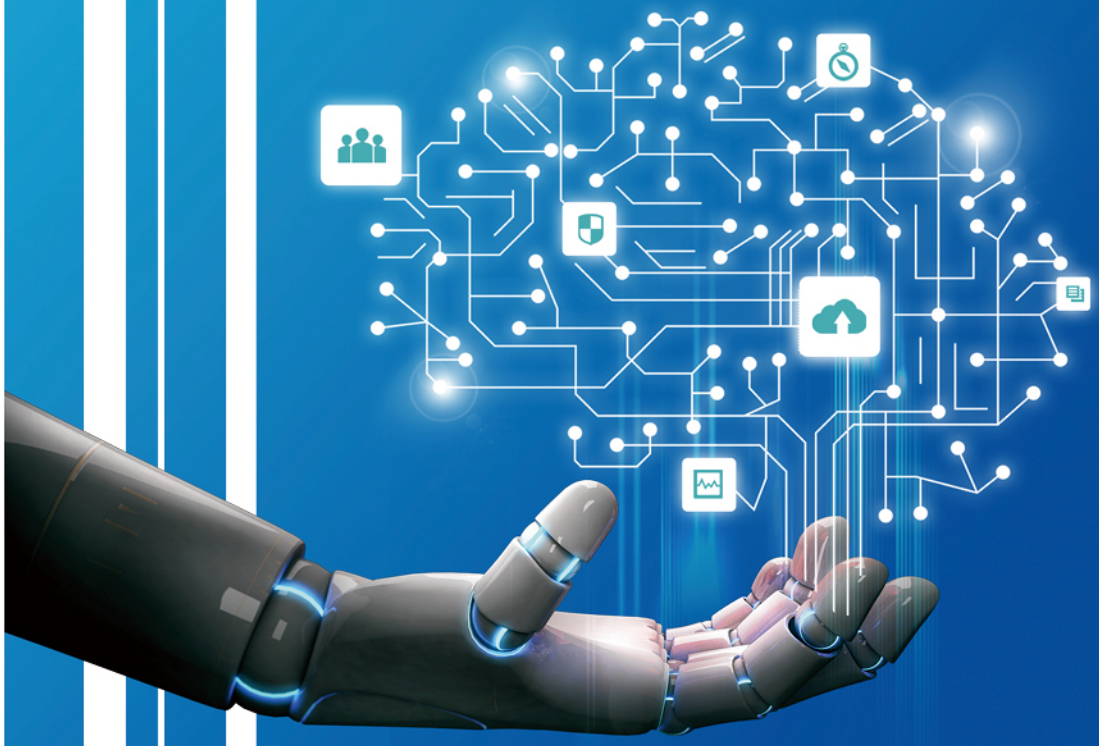
http://www.kiice.org

2019
4차 산업 관련
지능형 기술 인력 양성 워크숍

KIICE 한국정보통신학회 사단법인 한국정보통신학회

- 일시 : 2019년 8월 23일 금요일 9시 30분~17시
- 주최 : 한국정보통신학회 지능정보연구회
- 주관 : 한국정보통신학회, 부산인재평생교육진흥원
- 장소 : 신라대학교 국제관 442호



목차

환영사

Invitation

한국정보통신학회 회원 및 산업체 전문가 여러분,

현재 우리가 살고 있는 세상은 네트워크(IoT, 5G), 데이터(Cloud, Big Data), 인공지능(기계학습, 알고리즘) 등 지능 정보통신기술(ICT)이 기존 산업과 융합하여 새로운 서비스와 가치를 창출하는 4차 산업혁명 시대이며, 세상의 모든 것들이 빠르게 변화하는 새로운 사회의 도래를 목전에 두고 있습니다.

이러한 급격한 시대 변화의 중심에 한국정보통신학회가 자리하고 있습니다. 우리 학회는 22년이란 짧은 역사에도 불구하고 12회의 국제학술대회와 45회의 국내학술대회를 개최하였고 매달 발간되는 국문지와 분기별 발간되는 영문지가 한국연구재단의 등재지로서 그 권위를 더 하고 있습니다. 또한, 작년에는 영문논문지(JICCE)가 국제 유명인용 색인인 SCOPUS에 등재되는 성과를 이루었습니다.

우리 학회는 4차 산업혁명 시대를 맞이하여 IoT, 5G, AI, Big Data, 클라우드 등의 학문적 발전을 도모하고 관련 기술에 종사하는 학계 및 산업체와의 유대관계를 공고히 하고 이를 기반으로 우리나라 정보통신 산업을 선진국 수준으로 발전시켜야 하는 시대적 사명을 지고 있다고 생각합니다.

이에 한국정보통신학회 지능정보연구회 주관으로 개최되는 본 기술 워크숍은 매우 시기 절적하며, 관련 산업체에 계시는 분들 뿐만 아니라 이 분야를 연구하고 관심을 갖는 학생들에게도 인공지능 관련기술의 실무적 know-how와 깊이 있는 이론적 지식을 공유하는 의미 있는 자리가 될 것으로 생각됩니다. 금번 기술 워크숍에서의 활발한 의견교환과 지식공유를 통해 우리나라 ICT 산업이 한 단계 더 발전하는 기회가 되길 기대하며, 참석하신 모든 분들의 적극적인 참여를 부탁드립니다.

끝으로 본 기술 워크숍을 준비해주신 지능정보연구회 위원장이신 김광백교수님을 비롯한 관계자 분들의 노고에 깊은 감사를 드리며, 참석해 주신 모든 분들의 가정에 행복이 가득하시길 기원합니다. 고맙습니다.

2019년 8월

사단법인 한국정보통신학회 회장 **오창헌**

초대의 말씀

Invitation

4차 산업 혁명의 기반 기술은 인공지능 기술로서 기계 학습, 심층 신경망, 퍼지 시스템, 유전 알고리즘 등이 실무에서 다양하게 적용되고 있습니다. 최근에는 데이터과학(data science) 분야에 인공지능 기술이 중요시 되고 있습니다. 데이터 과학은 데이터마이닝(data mining)과 유사하게 정형, 비정형 형태를 포함한 다양한 데이터로부터 지식과 인사이트를 추출하는데 과학적 방법론, 프로세스, 알고리즘, 시스템을 동원하는 융합분야로서 데이터를 통해 실제 현상을 이해하고 분석하는데 통계학, 데이터 분석, 기계학습과 연관된 방법론을 통합하는 개념으로 정의되고 있습니다. 따라서 비정형 형태를 포한한 데이터로부터 지식과 인사이트를 추출하는 기법으로 인공지능 기술이 적용되고 있습니다.

그러나 대부분의 중소기업 현장에서는 인공지능 알고리즘 구조와 인공지능 알고리즘 장단점 분석 없이 오픈소스 등을 적용하고 있는 실정입니다. 따라서 경쟁력 있는 인공지능 서비스를 제공하지 못하고 있고 인공지능 서비스 평가 모델도 주관적으로 제시하므로 인공지능 서비스를 높이지 못하고 있는 실정입니다.

따라서 지능 정보 연구회의 2차 워크숍에서는 인공지능 서비스를 객관적으로 제공하기 위한 과정으로 퍼지 연상 메모리 기법, 심층 기반 퍼지 학습 알고리즘, 웹 스크래핑과 텍스트 마이닝, 의사결정트리를 활용한 분류 기법 등에 대해 강연하여 데이터 과학 분야에서 비정형 형태의 데이터를 객관적으로 분석할 수 있는 기반을 조성하고 인공지능 서비스를 높일 수 있도록 합니다.

본 워크숍은 현장에서 지능형 데이터 과학이나 빅데이터 처리에 기반 기술인 인공지능이 활용되는 분야에서 실질적으로 도움이 되도록 준비하였습니다.

이번 워크숍이 실무 개발자의 수준을 향상시키는 계기가 되기를 바라며, 개발자 및 연구자들의 적극적인 참여와 성원을 부탁드립니다.

감사합니다.

한국정보통신학회 지능정보연구회 위원장 **김광백**

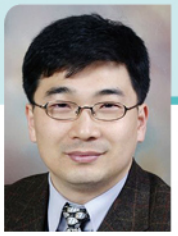
연사 소개

김광백 (Kwang Baek Kim) 교수
신라대학교

1999년 : 부산대학교 전자계산학과 졸업(이학박사)
1997년~현재 : 신라대학교 컴퓨터소프트웨어공학부 교수
2013년 : International Journal of Computational Vision and Robotics(SCOPUS), Guest Editor
2013년 : International Journal of Information and Communication Technology(SCOPUS), Guest Editor
2015년~2016년 : Computational Intelligence and Neuroscience(SCIE), Lead Guest Editor
2016년~2017년 : 한국정보통신학회 회장
2014년~현재 : Open Computer Science Journal, Editor.
2012년~현재 : 한국지능정보시스템학회 편집위원
2013년~현재 : International Journal of Intelligent Information Processing(SCOPUS), Editor
2015년~현재 : The International Journal of Fuzzy Logic and Intelligent Systems(SCOPUS), Associate Editor
2017년~현재 : IEEE Computational Intelligence Society 이사
논문(2013~2019) : SCIE(1저자 & 교신저자 : 29편), SCOPUS(1저자 & 교신저자 : 31편)
연구분야 : 심층 신경망, 기계 학습, 의료 영상 처리, 퍼지 시스템, 데이터마이닝

우영운(Young Woon Woo) 교수
동의대학교

1997년 8월 : 연세대학교 전자공학과 (공학박사)
1997년 9년~현재 : 동의대학교 응용소프트웨어공학과 교수
2016년~2018년 : 한국정보통신학회 국문지 편집위원장
2018년~현재 : 한국정보통신학회 수석부회장
2017년 5월 : 'AI 이해와 활용' 특강(KTDS)
2017년 8월 : '인공지능의 이해와 활용' 특강(인제대학교 대학원)
논문 및 학술발표(2013~2019) : SCOPUS(5편), 학진등재지(8편), 국제학술발표(6편), 국내학술발표(13편),
역서 〈엑셀로 배우는 인공지능〉 (제이펍, 2017)
연구 분야 : Artificial Intelligence, Machine Learning, Fuzzy Techniques, Pattern Recognition, Text Mining



김희철 (Hee-Cheol Kim) 교수
인제대학교

2001년 : Stockholm 대학교 수치해석/컴퓨터과학과 졸업(이학박사)
2002년~현재 : 인제대학교 컴퓨터공학부/헬스케어IT학과 교수
2016년~현재 : 인제대학교 일반대학원 디지털항노화헬스케어학과 학과장
2016년~현재 : 경상남도교육청 정보화정책심의위원
2017년~현재 : 한국정보통신학회 편집위원장
2005년~2010년 : IEEE HealthCom 편집위원
논문 및 연구 : HCI, 인공지능, 의료정보학 분야에서 100여 편 이상의 논문 출간.
〈나노섬유 기반 웰니스웨어 시스템 개발, 2009-2014〉,
〈창의융합산업 특성화 인재양성사업, 2016-2021〉 외 다수의 연구 및 교육 프로젝트를 연구책임자로 수행
저서 〈인간과 컴퓨터의 상호작용〉(사이텍미디어, 2006) : 문화관광부 우수학술도서
연구분야 : 인공지능, 기계학습, 빅데이터 마이닝, 디지털 헬스케어, 휴먼 컴퓨터 인터페이스

4차 산업 관련 지능형 기술 인력 양성 워크숍 프로그램

시간	주제	발표자
09:30 ~ 10:00	등록	
10:00 ~ 12:00	**1. 퍼지 연상 메모리와 심층 기반 퍼지 학습 방법** 1. 퍼지 연상 메모리 기법 2. 심층 퍼지 계층적 클러스터링 기법 3. Deep Fuzzy Supervised Learning Algorithm	**김광백 교수** (신라대학교)
12:00 ~ 13:00	점심시간 (점심 제공)	
13:00 ~ 14:50	**2. 웹 스크래핑과 텍스트 마이닝** 1. 온라인 뉴스 스크래핑 2. 토픽 모델링을 위한 전처리 3. 단어 빈도수 분석 및 워드 클라우드 4. LDA 기반의 토픽 추출 및 분석	**우영운 교수** (동의대학교)
14:50 ~ 15:00	Break	
15:00 ~ 16:50	**3. 의사결정트리를 활용한 분류기법** 1. 의사결정트리를 활용한 분류기법 2. 의사결정트리 활용 예와 랜덤포리스트 기법 3. 퍼지의사결정트리를 활용한 비선형 분류기법	**김희철 교수** (인제대학교)

퍼지 연상 메모리와 심층 기반 퍼지 학습 알고리즘

10:00~12:00

김광백 교수(신라대학교)

신라대학교

퍼지 연상 메모리와 심층 기반 퍼지 학습 알고리즘

Deep Neural Networks & Medical Imaging LAB
Silla University
Kwang Baek Kim

김광백 교수

신라대학교

1997년~현재 : 신라대학교 컴퓨터소프트웨어공학부 교수
2013년 : International Journal of Computational Vision and Robotics(SCOPUS),
2013년 : International Journal of Information and Communication Technology(SCOPUS), Guest Editor
2014년 : Neurocomputing(Elsevier :SCIE), Guest Editor
2015년~2016년 : Computational Intelligence and Neuroscience(SCIE), Lead Guest Editor
2016년 : Neural Computing and Applications (Springer : SCIE), Guest Editor
2016년 ~2017년 : 한국정보통신학회 회장
2014년~현재 : Open Computer Science Journal, Editor.
2012년~현재 : 한국지능정보시스템학회 편집위원
2015년~현재 : The International Journal of Fuzzy Logic and Intelligent Systems(SCOPUS), Associate Editor
2017년~현재 : IEEE Computational Intelligence Society 이사
논문(2013~2019) : SCIE(1저자 및 교신저자 : 29편), SCOPUS(1저자 및 교신저자 : 31편)
연구분야 : 심층 신경망, 기계 학습, 의료 영상 처리, 퍼지 시스템, 데이터마이닝

김광백 교수
신라대학교

신라대학교

1. 퍼지 연상 메모리

신라대학교

(1-1) Fuzzy Systems

Fuzzy 개념

신라대학교

퍼지 이론은 이치 논리가 아니라 다치 논리임
애매한 기준을 수치적으로 값을 표현

Example

뜨겁다 / 차갑다 (이치 논리)
-> 1, 0

매우 뜨겁다 / 뜨겁다 / 미지근하다 / 약간 차갑다 / 차갑다 (다치 논리)
-> 1, 0.8, 0.6, 0.4, 0.2, 0

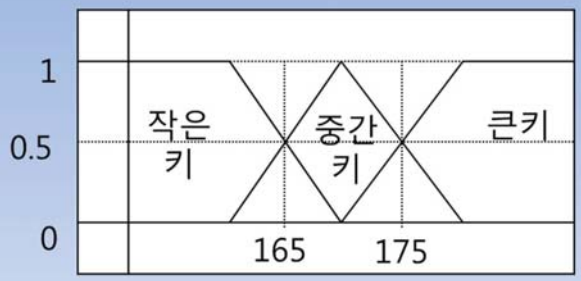
애매함을 수리적으로 취급이 가능하도록 하는 이론 → **퍼지이론**

5

Fuzzy Set

신라대학교

퍼지 집합



	큰키	중간키	작은키	결론
176cm	0.6	0.4	0	키가 큰 것보다 약간 작다
166cm	0	0.6	0.4	키가 중간 보다 약간 작다
163cm	0	0.3	0.7	키가 작다

6

합집합 증명

신라대학교

집합 A와 B가 전체집합 X의 부분집합들일 때 합집합 $A \cup B$의 특성 함수 $\chi_{A\cup B} : X \to \{0,1\}$는 임의 $x \in X$에 대해, 다음이 성립한다.

$$\chi_{A\cup B}(x) = \max\{\chi_A(x), \chi_B(x)\}$$

<증명> 특성 함수 $\chi_{A\cup B}$는 0과 1 두 값만을 취하므로 각각의 경우를 구분하여 증명한다.

(1) 원소 $x \in X$에 대해 $\chi_{A\cup B}(x) = 1$인 경우

$$\begin{aligned} \chi_{A\cup B}(x) = 1 &\Leftrightarrow x \in A \cup B \\ &\Leftrightarrow x \in A \text{ 또는 } x \in B \\ &\Leftrightarrow \chi_A(x) = 1 \text{ 또는 } \chi_B(x) = 1 \\ &\Leftrightarrow \max\{\chi_A(x), \chi_B(x)\} = 1 \end{aligned}$$

(2) 원소 $x \in X$에 대해 $\chi_{A\cup B}(x) = 0$인 경우

$$\begin{aligned} \chi_{A\cup B}(x) = 0 &\Leftrightarrow x \notin A \cup B \\ &\Leftrightarrow x \notin A \text{ 이고 } x \notin B \\ &\Leftrightarrow \chi_A(x) = 0 \text{ 이고 } \chi_B(x) = 0 \end{aligned}$$

따라서 (1)과 (2)로부터 모든 $x \in X$에 대해 $\chi_{A\cup B}(x) = \max\{\chi_A(x), \chi_B(x)\}$이 성립한다.

7

교집합 증명

신라대학교

집합 A와 B가 전체집합 X의 부분집합들일 때 합집합 $A \cap B$의 특성 함수 $\chi_{A\cap B} : X \to \{0,1\}$는 임의 $x \in X$에 대해, 다음이 성립한다.

$$\chi_{A\cap B}(x) = \min\{\chi_A(x), \chi_B(x)\}$$

<증명>

(1) 원소 $x \in X$에 대해 $\chi_{A\cap B}(x) = 1$인 경우

$$\begin{aligned} \chi_{A\cap B}(x) = 1 &\Leftrightarrow x \in A \cap B \\ &\Leftrightarrow x \in A \text{ 이고 } x \in B \\ &\Leftrightarrow \chi_A(x) = 1 \text{ 이고 } \chi_B(x) = 1 \\ &\Leftrightarrow \min\{\chi_A(x), \chi_B(x)\} = 1 \end{aligned}$$

(2) 원소 $x \in X$에 대해 $\chi_{A\cap B}(x) = 0$인 경우

$$\begin{aligned} \chi_{A\cap B}(x) = 0 &\Leftrightarrow x \notin A \cap B \\ &\Leftrightarrow x \notin A \text{ 또는 } x \notin B \\ &\Leftrightarrow \chi_A(x) = 0 \text{ 혹은 } \chi_B(x) = 0 \\ &\Leftrightarrow \min\{\chi_A(x), \chi_B(x)\} = 0 \end{aligned}$$

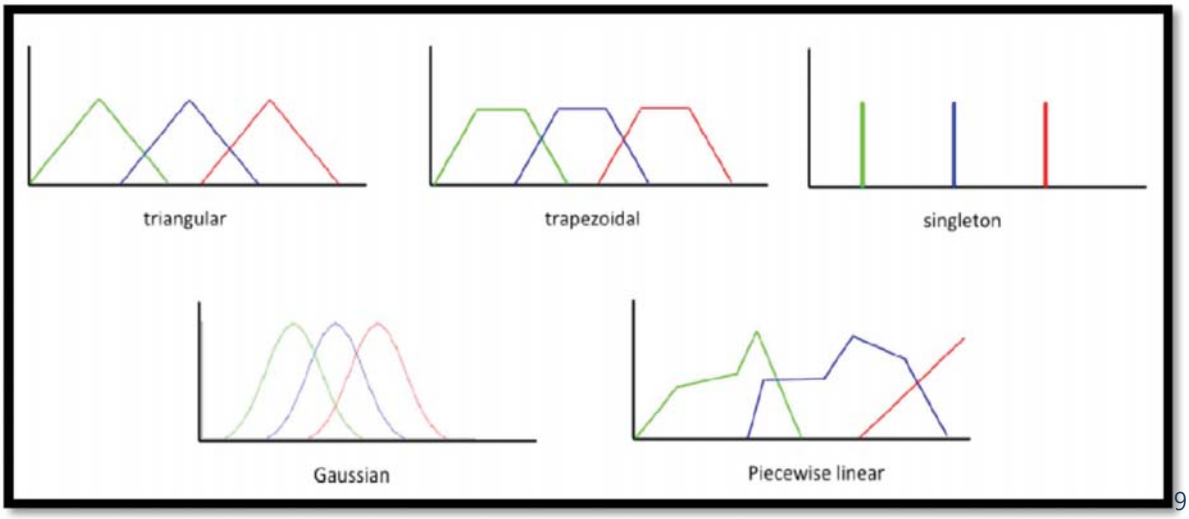
따라서 (1)과 (2)로부터 모든 $x \in X$에 대해 $\chi_{A\cap B}(x) = \min\{\chi_A(x), \chi_B(x)\}$이 성립한다.

8

Membership Functions

신라대학교

- Triangular.
- Trapezoidal.
- Piecewise linear.
- Gaussian.
- Singleton.



9

Fuzzy Inference

신라대학교

If x is A1 and y is B1, Then z is C1
If x is A2 and y is B2, Then z is C2

R1의 적합도 : $W_1 = \mu_{A1}(x_0) \wedge \mu_{B1}(y_0)$

R2의 적합도 : $W_2 = \mu_{A2}(x_0) \wedge \mu_{B2}(y_0)$

R1의 추론 결과 : $\mu_{c1}(z) = W_1 \wedge \mu_{c1}(z), \forall_z \in Z$

R2의 추론 결과 : $\mu_{c2}(z) = W_2 \wedge \mu_{c2}(z), \forall_z \in Z$

$$\mu_c(z) = \mu_{c1}(z) \vee \mu_{c2}(z)$$

$$z_0 = \frac{\int u_c(z) \bullet z dz}{\int u_c(z) dz}$$

10

신라대학교

(1-2) 연상메모리

연상 메모리의 분류

신라대학교

- **순방향 구조**
  - 선형 연상 메모리
- **순환 구조**
  - 순환 연상 메모리
  - Hopfield 모델
  - 양방향 연상 메모리

연상 메모리의 유형

신라대학교

- **이질 연상 메모리**

  입력 패턴과 연상되는 패턴이 다름

- **동질 연상 메모리**

  입력 패턴과 연상되는 패턴이 동일

Hopfield 모델

신라대학교

- 최종 출력을 얻을 때까지 외부 입력 활용
- 연결 강도 대칭 구조 / 대각 요소 0

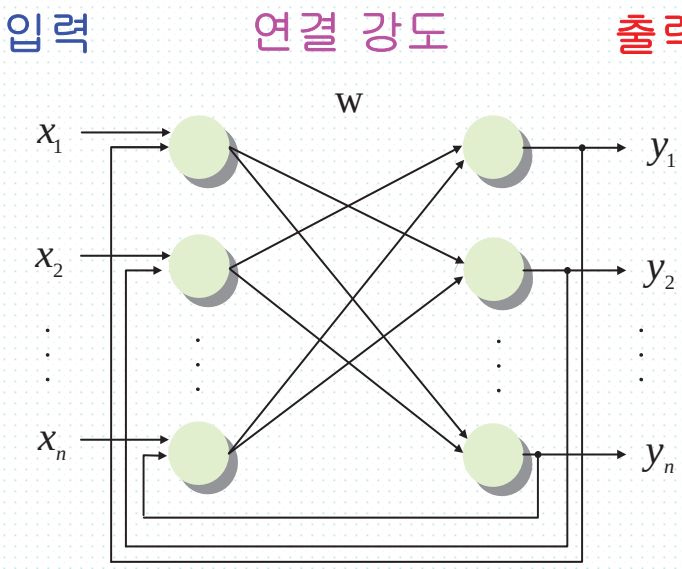
$$W_{ij} = W_{ji}$$

$$W_{ii} = 0$$

$$W = \sum_{i=1}^{p} s^T(i)s(i) - pI$$

Hopfield 모델의 구조

신라대학교



15

두 패턴 저장하는 Hopfield 모델

신라대학교

$$s(1) = [1 \;\; 1 \;\; -1 \;\; -1] \qquad s(2) = [-1 \;\; -1 \;\; 1 \;\; 1]$$

16

첫 번째 패턴 저장

신라대학교

$$s(1) = [1 \;\; 1 \;\; -1 \;\; -1]$$

$$W_1 = s(1)^T s(1) - I$$

$$= \begin{bmatrix} 1 \\ 1 \\ -1 \\ -1 \end{bmatrix} [1 \;\; 1 \;\; -1 \;\; -1] - I$$

$$= \begin{bmatrix} 1 & 1 & -1 & -1 \\ 1 & 1 & -1 & -1 \\ -1 & -1 & 1 & 1 \\ -1 & -1 & 1 & 1 \end{bmatrix} - \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 0 & 1 & -1 & -1 \\ 1 & 0 & -1 & -1 \\ -1 & -1 & 0 & 1 \\ -1 & -1 & 1 & 0 \end{bmatrix}$$

17

두 번째 패턴 저장

신라대학교

$$s(2) = [-1 \;\; -1 \;\; 1 \;\; 1]$$

$$W_2 = s(2)^T s(2) - I$$

$$= \begin{bmatrix} -1 \\ -1 \\ 1 \\ 1 \end{bmatrix} [-1 \;\; -1 \;\; 1 \;\; 1] - I = \begin{bmatrix} 0 & 1 & -1 & -1 \\ 1 & 0 & -1 & -1 \\ -1 & -1 & 0 & 1 \\ -1 & -1 & 1 & 0 \end{bmatrix}$$

18

두 패턴 저장하는 연결 강도

신라대학교

$$W = W_1 + W_2$$

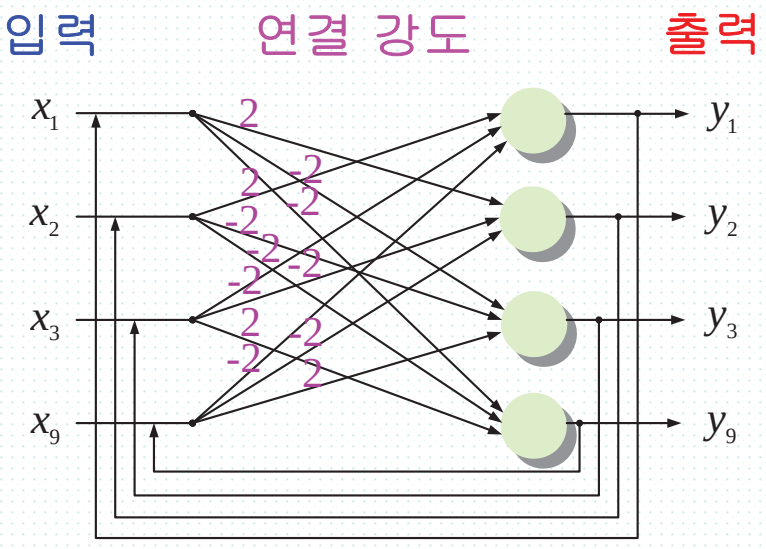
$$= \begin{bmatrix} 0 & 1 & -1 & -1 \\ 1 & 0 & -1 & -1 \\ -1 & -1 & 0 & 1 \\ -1 & -1 & 1 & 0 \end{bmatrix} + \begin{bmatrix} 0 & 1 & -1 & -1 \\ 1 & 0 & -1 & -1 \\ -1 & -1 & 0 & 1 \\ -1 & -1 & 1 & 0 \end{bmatrix}$$

$$= \begin{bmatrix} 0 & 2 & -2 & -2 \\ 2 & 0 & -2 & -2 \\ -2 & -2 & 0 & 2 \\ -2 & -2 & 2 & 0 \end{bmatrix}$$

19

Hopfield 모델 구현

신라대학교



20

Hopfield 모델 연상 알고리즘(1)

신라대학교

Step 1 : ***Compute weights to store P patterns***

$$W = \sum_{i=1}^{p} s^T(i)s(i) - p\mathrm{I}$$

Step 2 : ***Determine update order***

Step 3 : ***Set initial output***

$$Y \leftarrow X$$

Step 4 : ***For each unit*** y_i

Do Step 5-7

Step 5 : ***Compute NET***

$$NET = x_i + YW^T$$

21

Hopfield 모델 연상 알고리즘(2)

신라대학교

Step 6 : ***Update intermediate output***

$$y_i = \begin{cases} 1 : NET > 0 \\ y_i : NET = 0 \\ -1 : NET < 0 \end{cases}$$

Step 7 : Test stop condition

If y is converged, stop

else, change i according to predetermined order

and goto Step 4

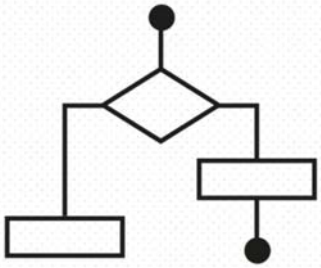
22

신라대학교

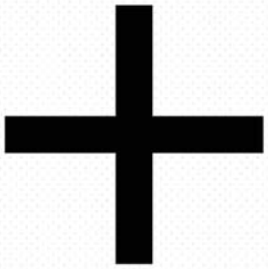
(1-3) 퍼지 연상 메모리

퍼지 연상 메모리

신라대학교



Associative
Memory
Algorithm



Fuzzy
Theory

24

Fuzzy Associative Memory

신라대학교

$S \circ W = S'$

[출력 값 계산]

$W = S^{T} \Lambda S$

[연결 강도]

° 연산은 Max-Min 합성 연산자

연결 강도 W, 입력 패턴 S

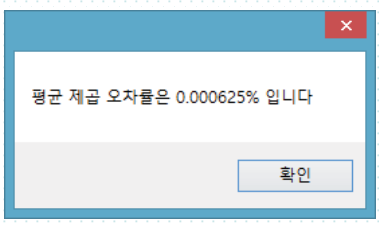
S ° W = S′ if and only if height(S) ≥ height(S′)

[패턴 복원]

25

MSE (평균 제곱 오차(Mean Squared Error) 신라대학교

$$MSE = \frac{1}{n \times m} \sum_{i=0}^{n-1} \sum_{j=0}^{m-1} (x_{ij} - \widehat{x_{ij}})$$



26

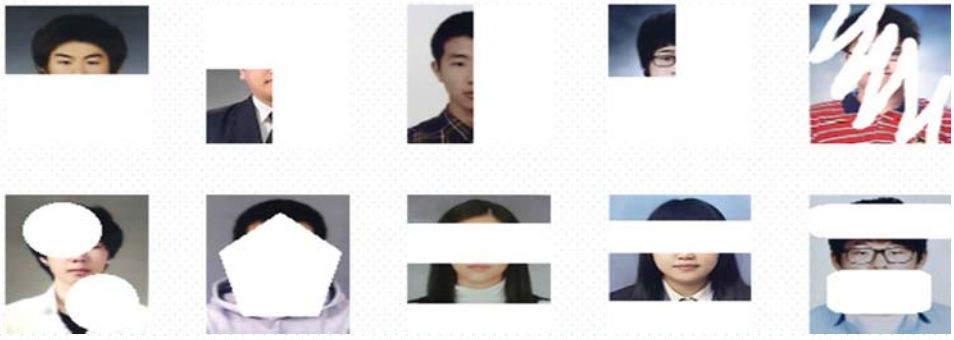
퍼지 연상 메모리 실험

신라대학교

SILLA university 1954

□ 실험 영상

□ 손상된 영상



27

퍼지 연상 메모리 실험

신라대학교

SILLA university 1954

□ 실험 영상

□ 손상된 영상

28

퍼지 연상 메모리 복원 결과

신라대학교

□복원 영상 & 오차값

29

신라대학교

2. 심층 기반 퍼지 클러스터링 기법

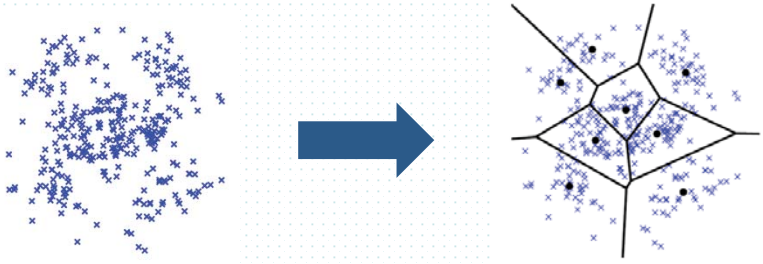
신라대학교

(2-1) 퍼지 클러스터링

Clustering

신라대학교

- **Clustering**
  - Unsupervised 학습
  - 유사성의 개념을 바탕으로 데이터를 몇 개의 그룹으로 분류하는 방법
  - 패턴인식, 문헌검색 등에 널리 응용되고 있음
  - ❖ Cluster : 사전적으로 무리라는 의미로 전산 분야에서는 동일 속성을 갖는 대상들을 하나로 묶은 대상을 의미

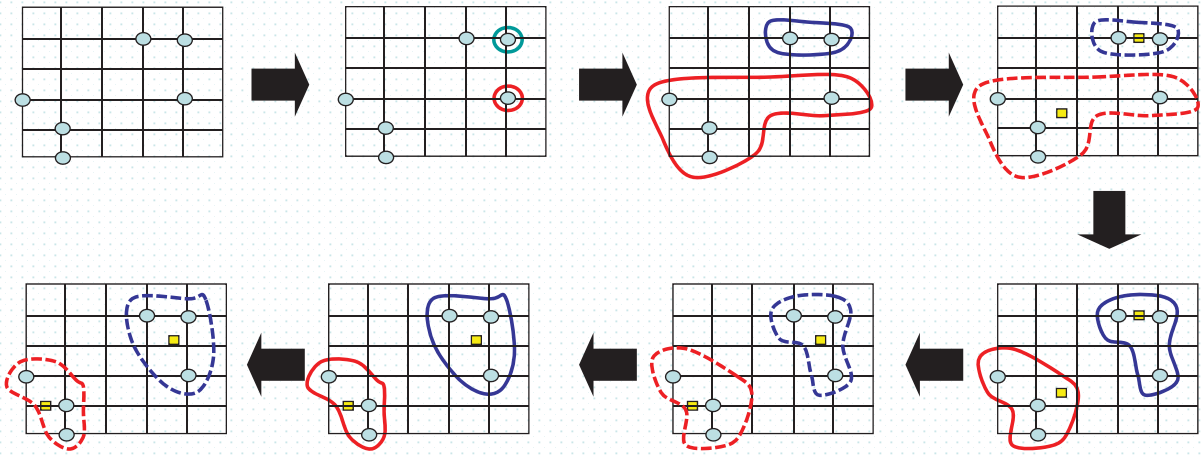


32

Clustering 과정

신라대학교

- **Clustering Algorithm**



33

클러스터링 기법 비교

신라대학교

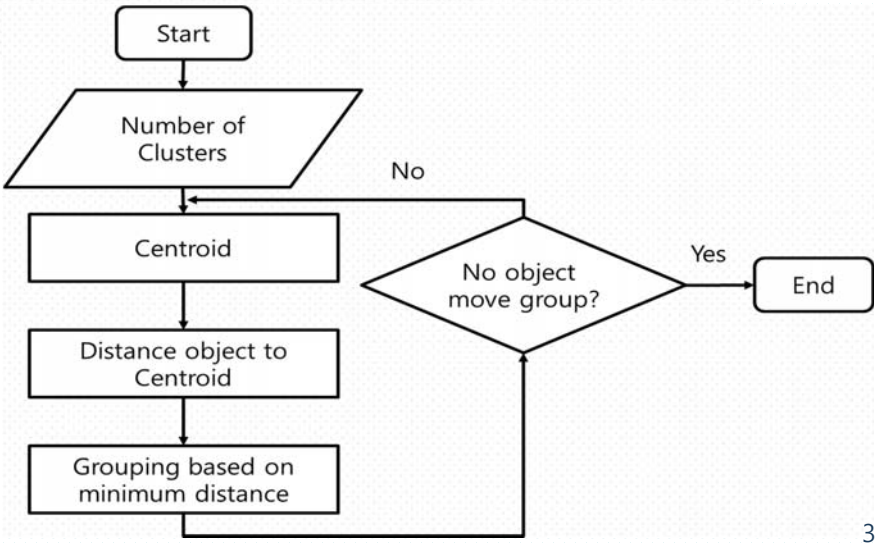
알고리즘	데이터 특징	클러스터 생성	Deep Unsupervised Neural Network 확장
K-Means	정량화된 빅데이터	정적 생성	가능 (효율적)
	비정량화된 빅데이터		불가능
FCM	정량화된 빅데이터	정적 생성	가능 (효율적)
	비정량화된 빅데이터		가능 (효율적)

34

K-Means 알고리즘

신라대학교

- K-Means 알고리즘의 개념은 패턴들과 그 패턴이 속하는 클러스터의 중심과의 평균 유클리디안(Euclidean)거리를 최소화 하는 것이다.
- K-Means 알고리즘의 성능은 초기 중심을 어떻게 선정하는가에 따라 크게 달라진다.



35

FCM(Fuzzy C-Means)

신라대학교

- **FCM(Fuzzy C-Means)**
  - 각각의 데이터를 각 클러스터에 속하는 소속 정도를 이용하여 데이터 분류하는 알고리즘
  - 유사성을 유클리디언 거리값을 이용하여 측정
  - HCM 알고리즘은 하나의 데이터가 하나의 클러스터에만 속할 수 있지만 FCM은 퍼지 소속정도를 이용하여 2개 이상의 클러스터에도 속할 수 있음

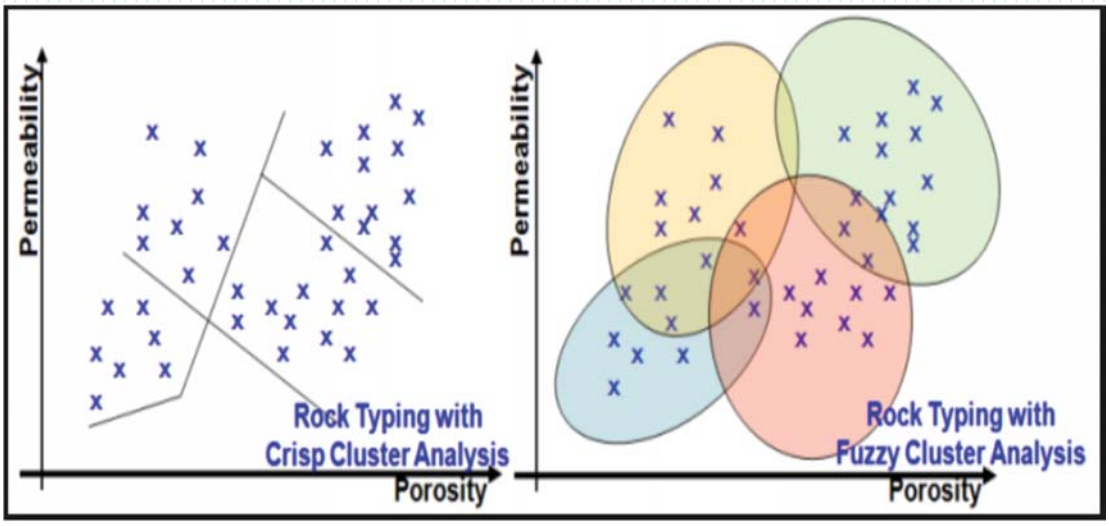
❖ **EM알고리즘** : Expectation과 Maximization단계를 반복하여 최적해를 찾는 알고리즘

❖ **HCM(K-Means)** : 유클리디안 거리를 이용하여 K개의 그룹으로 분류하는 클러스터링 알고리즘

36

Crisp Clustering & Fuzzy Clustering

신라대학교



FCM(Fuzzy C-Means)

신라대학교

- **FCM(Fuzzy C-Means) Algorithm**
  1. 소속함수 U를 초기화하고 파라메터 설정

$$\sum_{i=1}^{c} u_{ik} = 1, \qquad \forall k = 1, \cdots, n$$

$$0 < \sum_{k=1}^{n} u_{ik} < n$$

c : $2 \le c < n$(패턴의 개수)

m : $m \ge 1$

HCM

$$U = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \end{bmatrix}$$

FCM

$$U = \begin{bmatrix} 0.5 & 0 & 0 & 0.2 & 0 \\ 0.5 & 0.7 & 0 & 0.8 & 0.1 \\ 0 & 0.3 & 1 & 0 & 0.9 \end{bmatrix}$$

FCM(Fuzzy C-Means)

신라대학교

- **FCM(Fuzzy C-Means) Algorithm**

  2. 클러스터의 중심 계산

$$V_{ij} = \frac{\sum_{k=1}^{n} (u_{ik})^m x_{kj}}{\sum_{k=1}^{n} (u_{ik})^m}$$

39

FCM(Fuzzy C-Means)

신라대학교

- **FCM(Fuzzy C-Means) Algorithm**

  3. 소속함수 갱신

$$d_{ik} = d(x_k - v_i) = \|x_k - v_i\| = \left[\sum_{j=1}^{p} (x_{kj} - v_{ij})^2 \right]^{1/2}$$

HCM : $u_{ik}^{(r+1)} = 1 \quad , \min(d_{ik}^r) \quad j \in c$

$u_{ik}^{(r+1)} = 0 \quad , Otherwise$

FCM : $u_{ik}^{(r+1)} = \frac{1}{\sum_{j=1}^{c} (\frac{d_{ik}^r}{d_{jk}^r})^{2/m-1}}$

but $u_{ik}^{(r+1)} = 0 \quad , \quad I_k = \Phi$

40

FCM(Fuzzy C-Means)

신라대학교

- **FCM(Fuzzy C-Means) Algorithm**
  4. 종료 조건

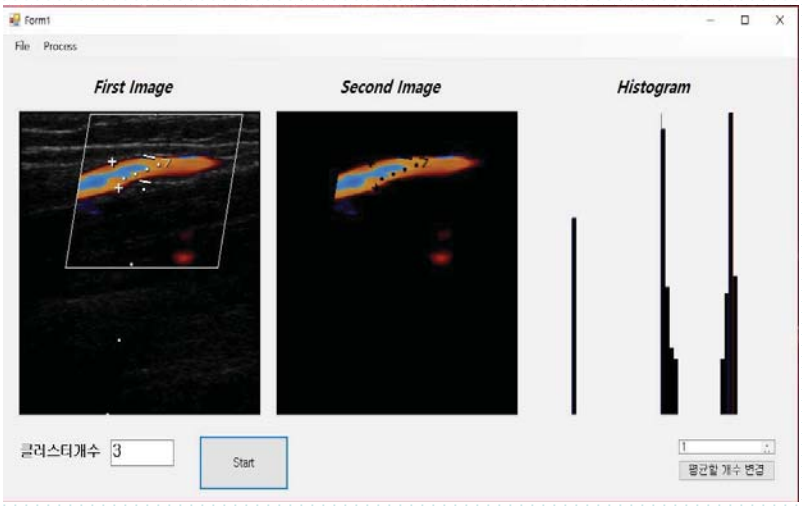
$$if\ (\max\left\|U^{(r+1)} - U^{(r)}\right\| < \varepsilon)\quad STOP$$

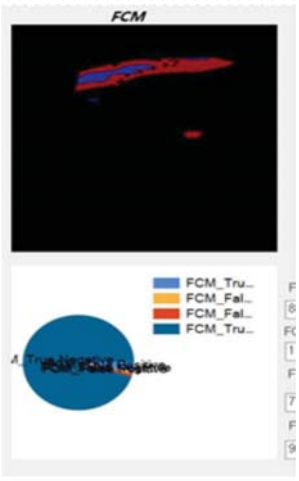
$$ELSE\ \text{단계}\,2\text{로}\ldots$$

41

FCM 기반 양자화

신라대학교





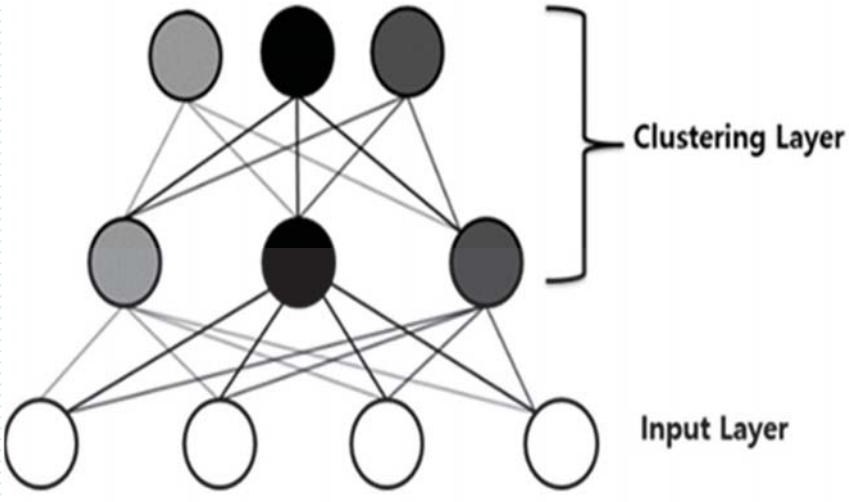
42

신라대학교

(2-2) 심층 기반 FCM

Deep FCM Model

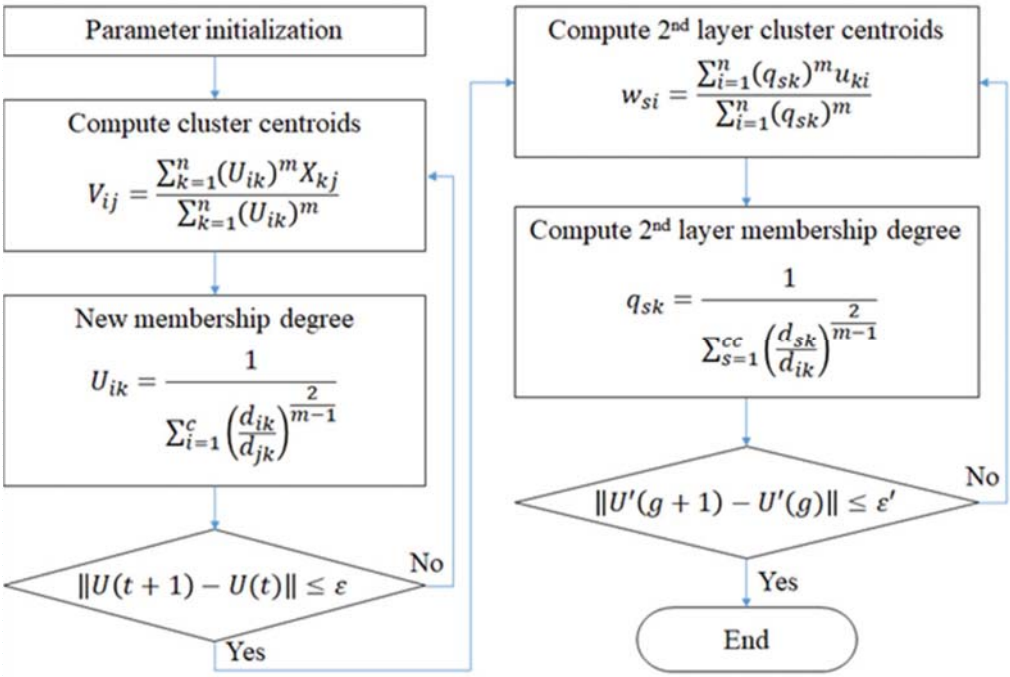
신라대학교



44

Deep FCM Algorithm

신라대학교



45

Deep FCM 기반 충수염

신라대학교

충수란?

6~9cm 길이의 근육질의 좁은 관으로 한쪽 끝은 막혔고, 다른 쪽 끝은 맹장과 붙어 있으며 대장이 시작되는 첫 부분으로 가느다란 관모양의 관

충수염이란?

충수 돌기가 어떤 원인에 의해 입구가 막히거나 부어있는 경우

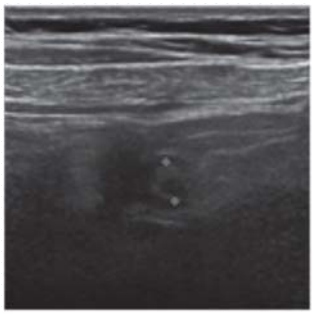
초음파검사의 문제점

검사자의 숙련도와 피로도에 따라 진단 결과가 달라질 수 있음

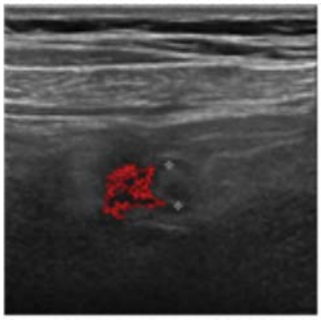
46

Deep FCM 기반 충수염

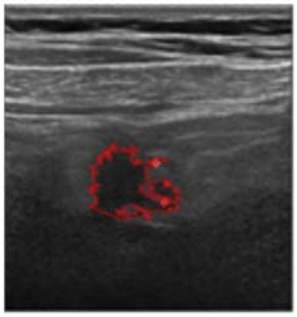
신라대학교



충수염 초음파 영상



FCM 결과



Deep FCM 결과

47

신라대학교

3. 심층 퍼지 학습 알고리즘

신라대학교

(3-1) Fuzzy Max-Min Neural Networks

지도 학습 Supervised Learning

신라대학교

지도 학습

분류
Classification

50

비지도 학습 Unsupervised Learning

신라대학교

비지도 학습

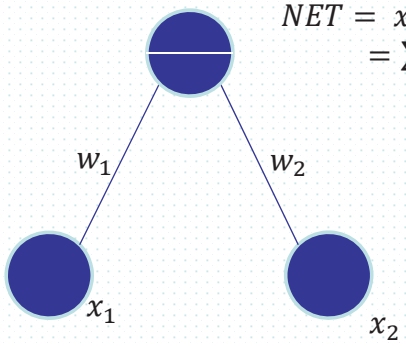
군집화
Clustering

51

Fuzzy Max-Min Neural Networks

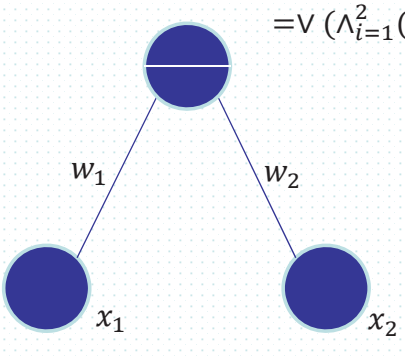
신라대학교

$if\ NET\ \geq 1\ Then\ Y = 1$
$else\ Y = 0$
$NET = x_1 w_1 + x_2 w_2$
$= \sum_{i=1}^{2} x_i w_i$



단층 퍼셉트론

$Y = NET \vee \theta_i$
$NET = (x_1 \wedge w_1) \vee (x_2 \wedge w_2)$
$= \vee (\wedge_{i=1}^{2}(x_i, w_i))$

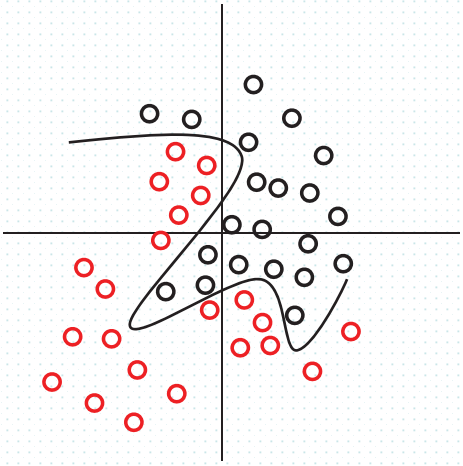


Fuzzy Max-Min Neural Networks

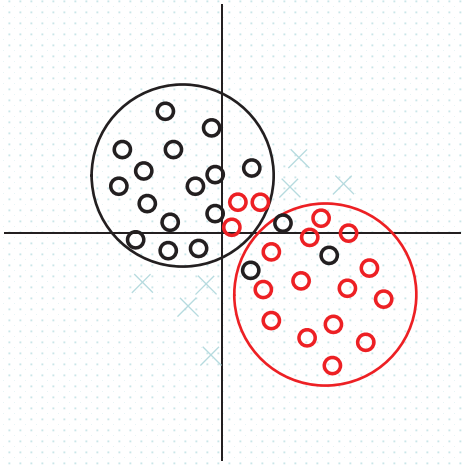
52

지도 학습 & 비지도 학습

신라대학교



지도 학습



비지도 학습

53

학습단계 분석

신라대학교

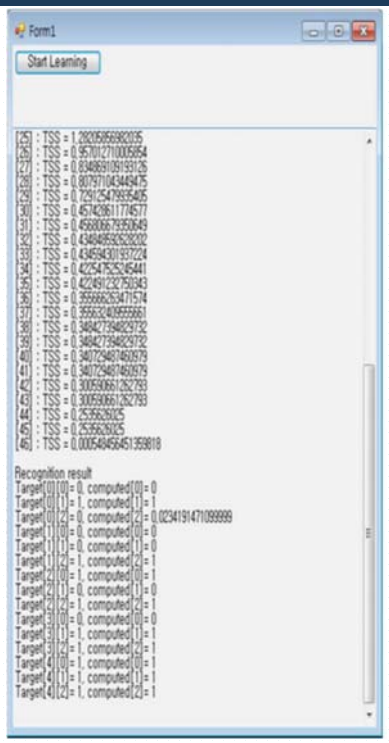
학습 단계

- 지도 학습 알고리즘

1. **수렴 단계**
   -학습 초기에 오차가 급격히 감소하는 단계

2. **경쟁 단계**
   -오차의 변화가 거의 없거나 오차 값이 진동하는 단계

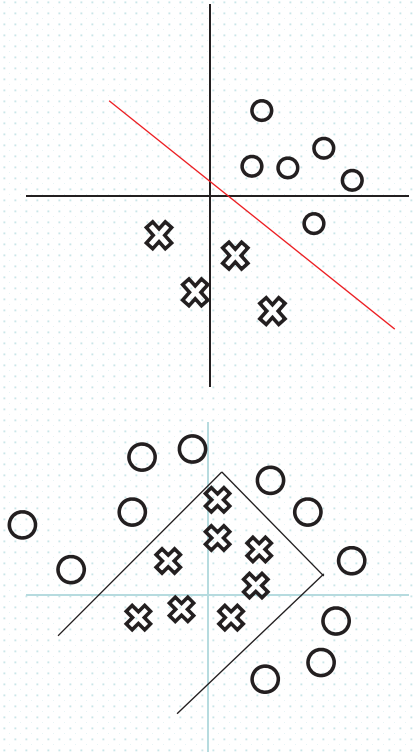
3. **우세 단계**
   -경쟁 단계를 벗어나 오차가 급격히 감소하면서 학습되는 단계

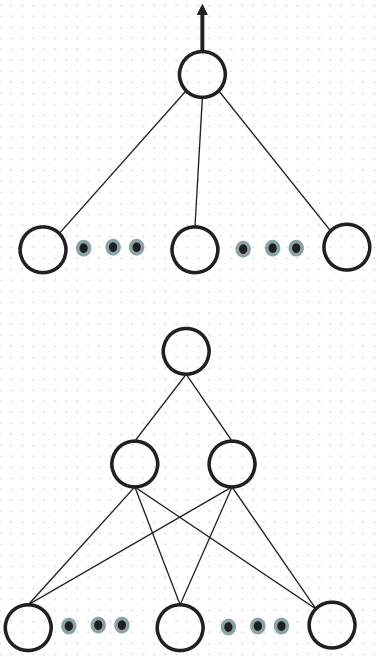


54

학습 구조 기반 분류 방법

신라대학교





55

퍼지 Max-Min 신경망 알고리즘

신라대학교

- NET는 입력 벡터(x_i)와 연결 가중치(w_i)의 최대-최소 합성 연산에 의해 식(1)과 같이 계산

$$NET = \vee_{j=1}^{q} \{\wedge_{i=1}^{p} (x_i, w_{ji})\} \quad \text{식(1)}$$

- 출력 벡터(o_j)는 식(2)와 같이 계산

$$o_j = NET \vee \theta_j \quad \text{식(2)}$$

56

퍼지 Max-Min 신경망 알고리즘 신라대학교

- 출력 벡터(o_j)와 목표 벡터(t_j)가 동일한 경우에는 연결가중치(w_i)와 바이어스항(θ_j)을 변경하지 않음
- 동일하지 않을 경우 식(3)에 의해 연결가중치(w_i)와 바이어스항(θ_j)을 변경

$$\Delta w_{ji}(n) = \Delta w_{ji}(n-1) + \frac{\partial o_i}{\partial w_{ji}}(t_j - o_j) \qquad \frac{\partial o_j}{\partial w_{ji}} = 1, when\ o_j = w_{ji},$$

$$= 0\ otherwise.$$

$$\Delta\theta_j(n) = \Delta\theta_j(n-1) + \frac{\partial o_j}{\partial \theta_j}(t_j - o_j) \qquad \frac{\partial o_j}{\partial \theta_j} = 1, when\ o_j = \theta_j,$$

$$= 0\ otherwise.$$

$$w_{ji}(n+1) = w_{ji}(n) + \alpha\Delta w_{ji}(n) + \beta\Delta w_{ji}(n-1)$$

$$\theta_j(n+1) = \theta_j(n) + \alpha\Delta\theta_j(n) + \beta\theta_j(n-1)$$

식(3)

57

퍼지 Max-Min 신경망 알고리즘 신라대학교

- 종료조건
  총 오차 자승합(TSS)이 오류 한계(ρ)보다 크면 단계 3으로 가고
  오류 한계보다 적거나 같으면 학습을 종료

$$TSS = \sum_{p=1}^{h}\sum_{j=1}^{q}\frac{1}{2}(t_j{}^p - o_j{}^p)^2$$

$$TSS > E_{MAX}\ Then\ Step\ 3.$$

$$otherwise\ 학습종료$$

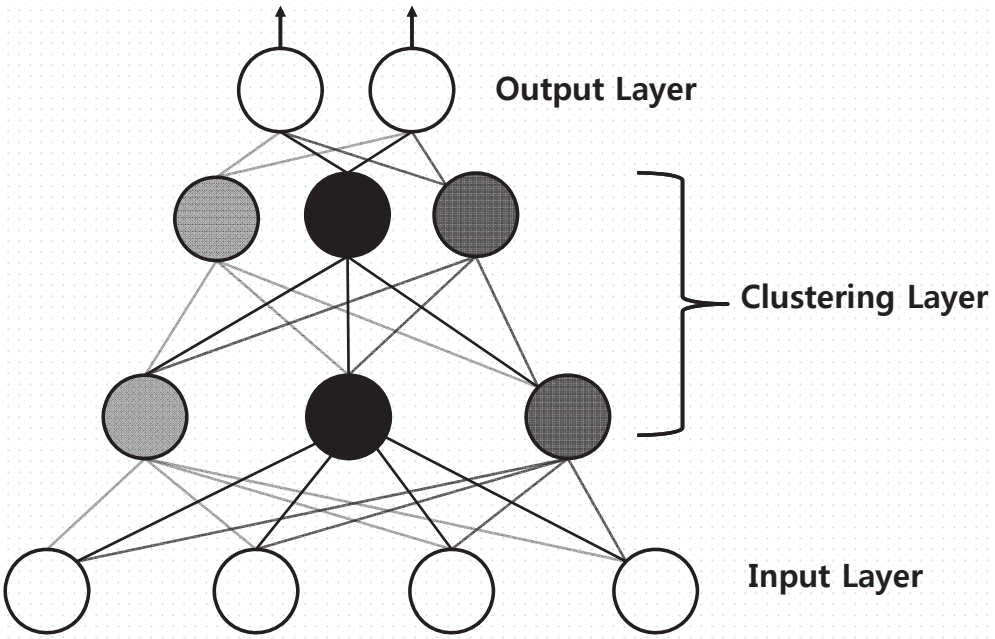
58

신라대학교

(3-2) Deep Fuzzy Learning Algorithm

Deep Fuzzy Learning Algorithm

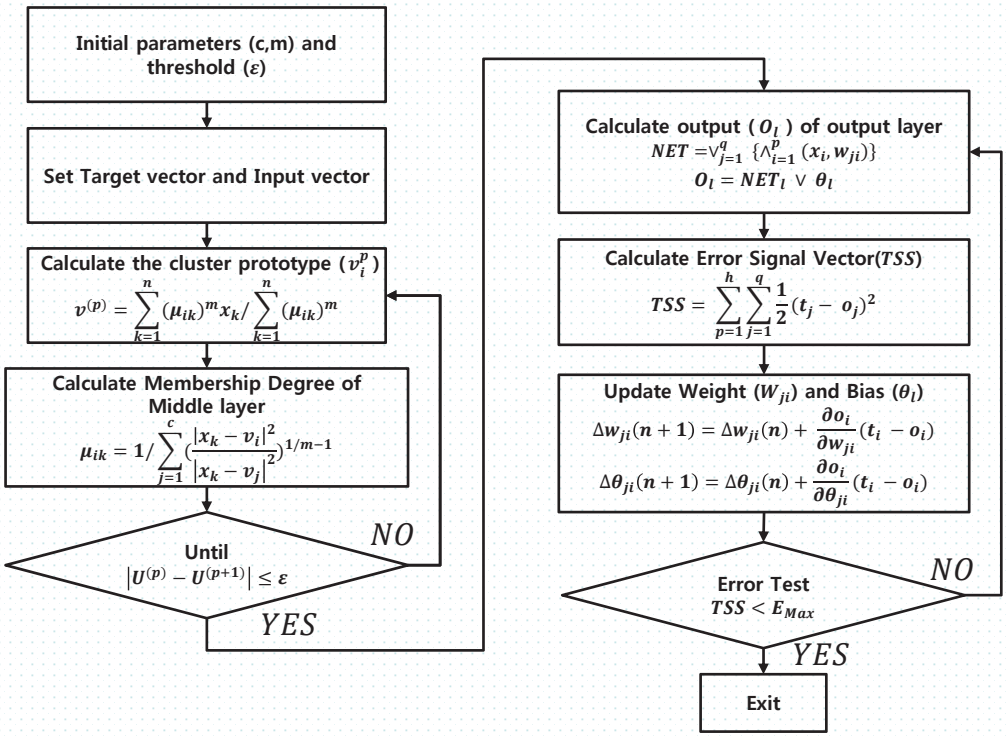
신라대학교



60

Deep Fuzzy Learning Algorithm

신라대학교



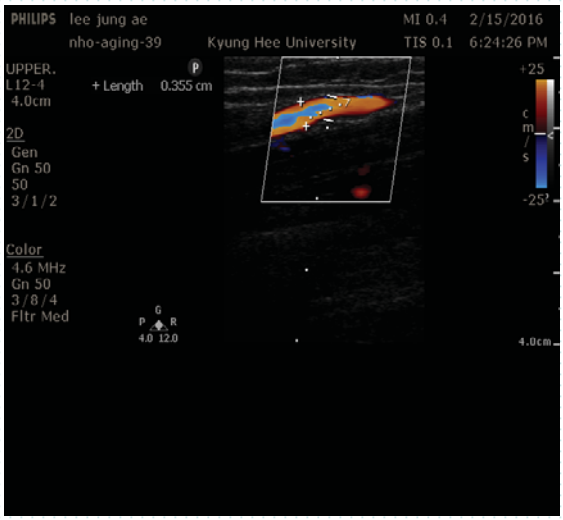
61

신라대학교

4. IHb 색상 정보를 이용한 색조 도플러 초음파 영상에서 상완 동맥의 동맥 혈류 속도 분석

연구 배경

신라대학교



색조 도플러 초음파 영상

- **상완 동맥이란?**
- **색조 도플러 초음파 영상이란?**
- **색조 도플러 초음파 영상 진단의 문제점**

63

색조 도플러 초음파 영상에서 동맥 영역 추출 과정

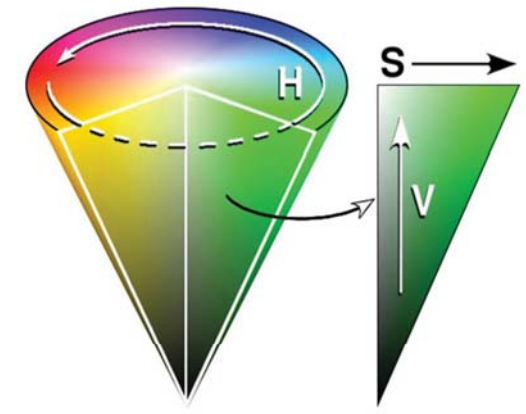
신라대학교

색조 도플러 초음파 영상
↓
ROI 영역 추출 및 이진화
↓
4방향 윤곽선 추적 기법을 이용한 사다리꼴 형태 추출
↓
고혈압 영역 명암도 강조
→
새로운 무게중심 기반 양자화
↓
동맥 영역 추출
↓
IHb기반 혈류 속도 분석

64

RGB 채널 HSV 채널 변환

신라대학교



HSV 색상표

$$V = \max(R, G, B)$$

$$S = \begin{cases} \frac{V - MIN(R,G,B)}{V} & if\ V \neq 0 \\ 0 & if\ V = 0 \end{cases}$$

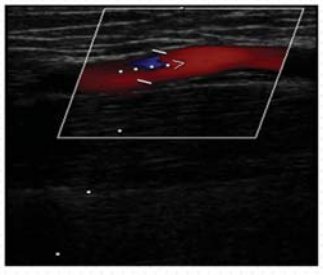
$$H = \begin{cases} \frac{60(G-B)}{V - min(R,G,B)} & if\ V = R \\ 120 + \frac{60(B-R)}{V - min(R,GB)} & if\ V = G \\ 240 + \frac{60(R-G)}{V - \min(R,G,B)} & if\ V = B \end{cases}$$

If H < 0 H = H + 360

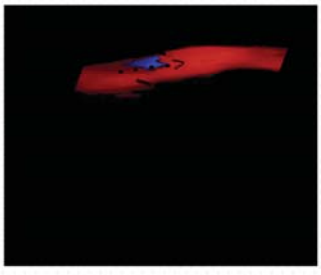
상완 동맥 영역 강조

신라대학교

If = $V_r * \left(1 + (1.0 - V_r)\right) > 1.0$
than $V_{New} = 1.0$
Else
than $V_{New} = V_r * (1 + (1.0 - V_r))$



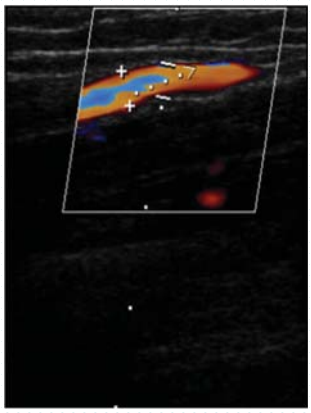
원본영상

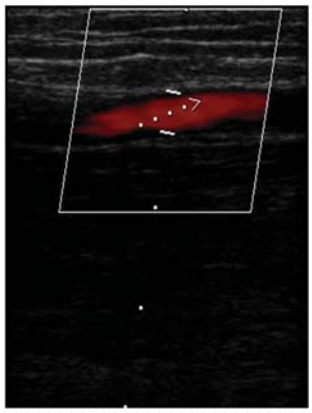


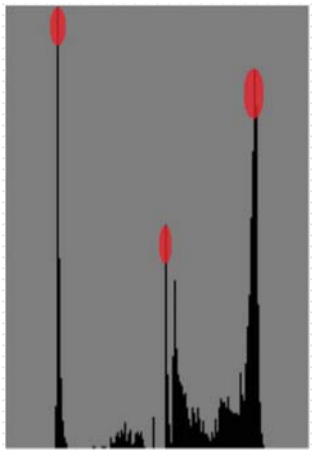
상완 동맥 영역 강조 결과

클러스터 개수 정의

신라대학교







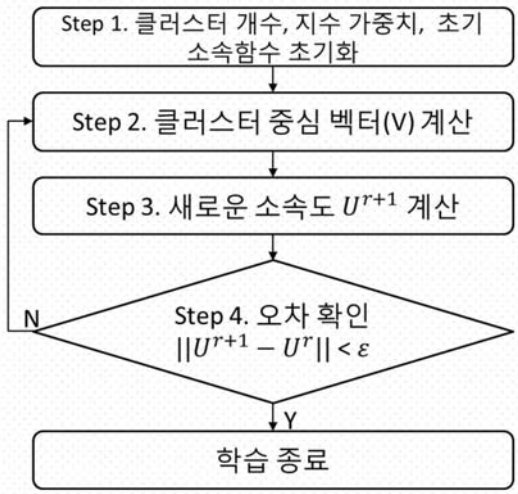
히스토그램 분석을 통한
봉우리 지점 개수 추출

67

무게 중심 계산

신라대학교

1. Fuzzy C_Means 알고리즘



식 1. $u_{ij} = \frac{\sum_{k=1}^{n}(u_{ik})^m x_{kj}}{\sum_{k=1}^{n}(u_{ij})^m}$

식 2. $d_{ik} = \left[\sum_{j=1}^{l}(x_{kj} - u_{ij})^2\right]^{1/2}$

식 3. $u^{(r+1)_{ik}} = \frac{1}{\sum_{j=1}^{c}\left[\frac{d_{ik}^r}{d_{jk}^r}\right]^2}$ for $I_{K \neq \emptyset}$

식 4. $||U^{r+1} - U^r|| < \varepsilon$ N

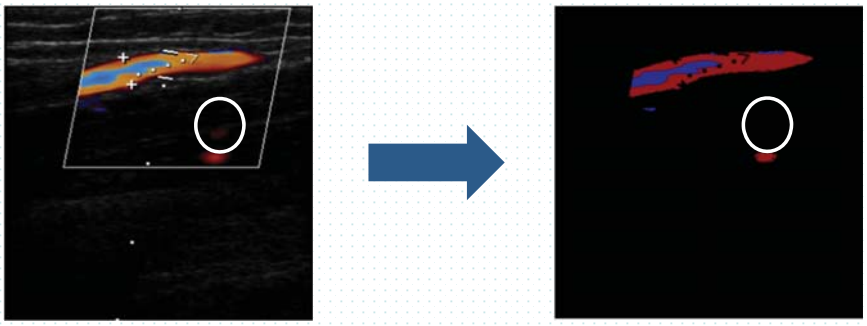
Y

학습 종료

68

FCM 양자화 결과

신라대학교

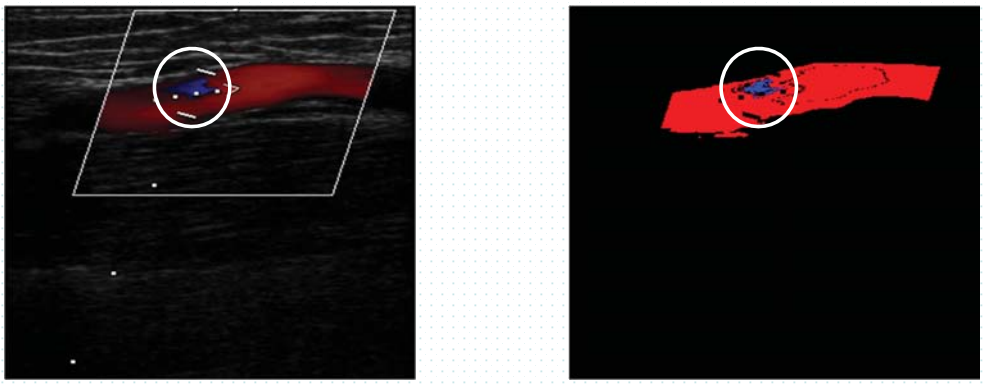


원본 영상 FCM 결과

69

제안된 무게중심 기반 양자화

신라대학교



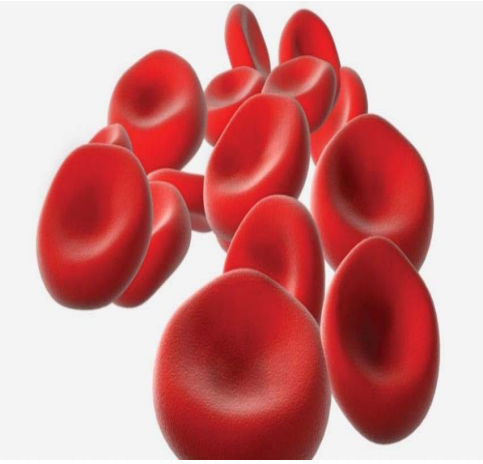
원본영상 FCM 기반 양자화 결과

70

IHb 기반 혈류 속도 분석

신라대학교

혈색소 지수 (IHb(Index of Hemoglobin)) 란?

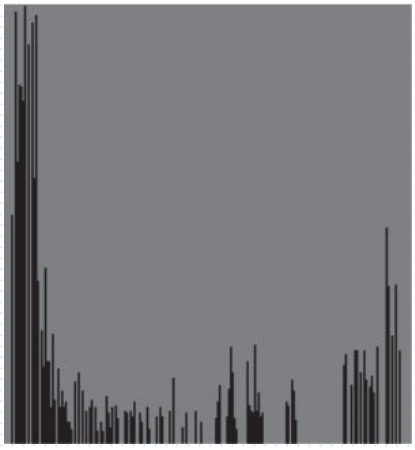


Step #1. $IHb = 32 * \log_2 \frac{V_r}{V_g}$

71

IHb 기반 혈류 속도 분석

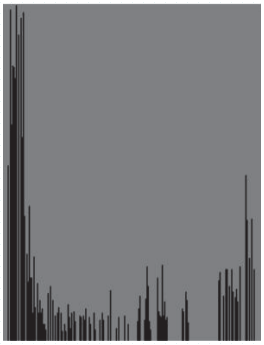
신라대학교



72

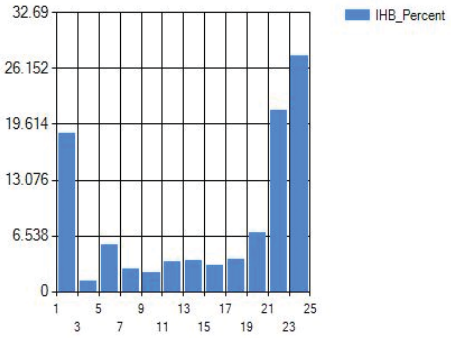
IHb 기반 혈류 속도 분석

신라대학교



IHb 히스토그램화

색상표 히스토그램화



IHb기반 혈류 속도 분석 결과

73

웹 스크래핑과 텍스트 마이닝

13:00~14:50

우영운 교수(동의대학교)

웹 스크래핑과 텍스트 마이닝

동의대학교 응용소프트웨어공학전공 우영운

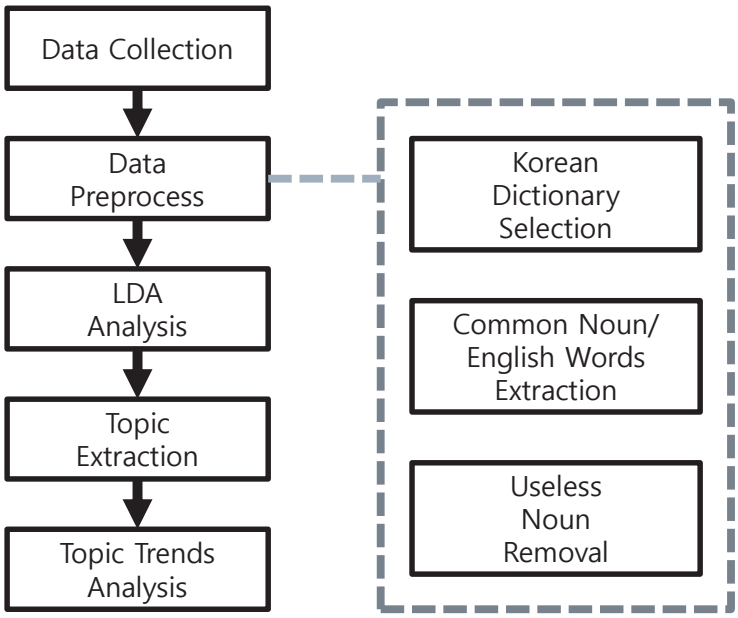
ywwoo@deu.ac.kr

♣ 웹 스크래핑과 텍스트 마이닝 연구 수행 사례

- 최근 웹 스크래핑 기법의 발달에 따라 대규모 웹 문서들을 자동으로 수집하고 분석하는 텍스트 마이닝 기법이 활발히 사용
- 텍스트 마이닝을 구현하는 오픈 소스 라이브러리가 풍부하게 제공됨으로 적은 분량의 코딩으로 대규모 데이터의 분석이 가능
- 이 논문에서는 2007년 1월부터 2016년 12월까지의 10년 동안 게재된 **JKIICE(한국정보통신학회 논문지)**에 게재된 논문지들의 대표 연구 주제와 연구 주제들의 년도별 변화 추이를 분석

♣ 웹 스크래핑과 텍스트 마이닝 연구 수행 사례

1) 전체 처리 과정



♣ 웹 스크래핑과 텍스트 마이닝 연구 수행 사례

2) 자료 수집

- 2007년부터 2016년까지 10년 동안 JKIICE에 게재된 논문 3,668편의 **'발간년월', '국문제목', '국문초록', '한글키워드'** 를 논문 별로 수집
- 이를 위해 Python 웹 스크래핑 프로그램을 작성하여 한국연구재단의 KCI 통합검색 사이트에서 자동으로 수집한 후 CSV 형태로 저장
- 수집된 4가지 정보들 중 **'국문제목', '국문초록', '한글키워드'** 를 하나의 문자열로 통합하여 **'발간년월', '한글논문데이터'**의 2개 필드의 데이터로 변환
- 최종적으로 변환된 3,668개의 데이터 쌍을 분석에 활용

♣ 웹 스크래핑과 텍스트 마이닝 연구 수행 사례

3) 자료 전처리

- 자료 전처리 및 분석을 위해 **RStudio와 R 언어**를 사용
- 한글 형태소 분석 후 보통 명사와 영어 단어 추출을 위해 R 언어에서 제공되는 **NIADic 사전**과 **SimplePos22() 함수** 활용
- 국문 초록이라 하더라도 전문 용어인 경우에는 영어 단어를 직접 기술하는 경우가 많아 영어 단어도 추출하여 주제 분석에 활용
- 추출된 보통 명사와 영어 단어들 중에 빈도수가 높으나 연구 주제와 관련이 없는 단어(예, *논문, 제안, 방법, 연구, 기반, 이용, 결과, 기법 ,각종, the, system, ieee* 등)들을 제거한 후, 남는 단어들 만을 논문 별로 다시 저장하여 전처리 과정 완료

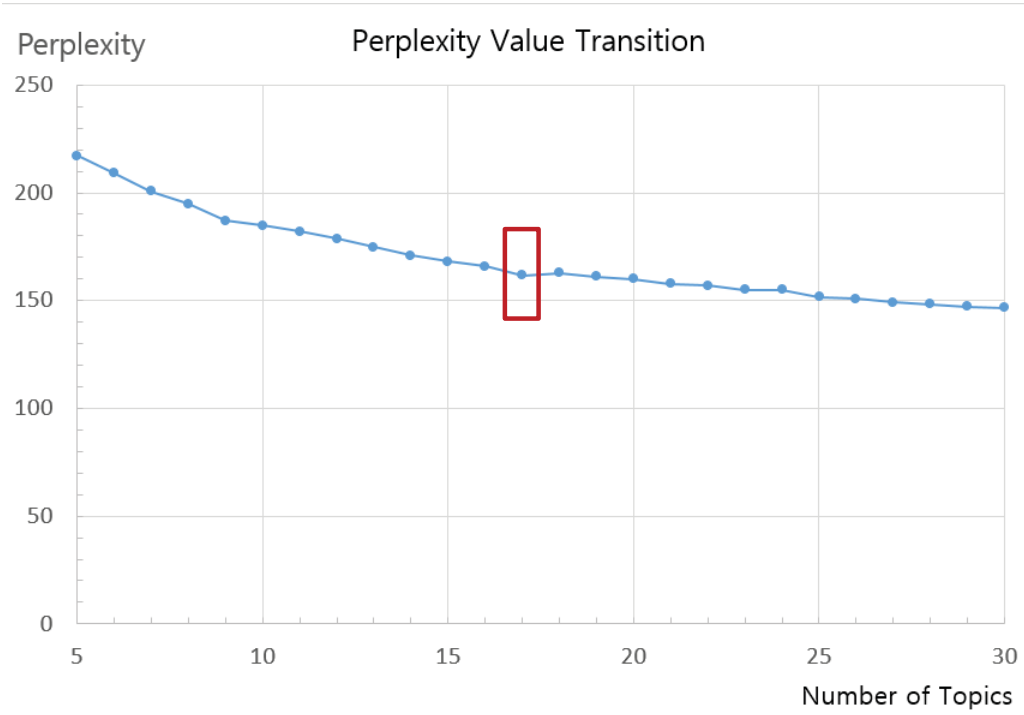
♣ 웹 스크래핑과 텍스트 마이닝 연구 수행 사례

4) 자료 분석

- 토픽 모델링에 의한 주제 분석을 위해 **LDA(Latent Dirichlet Analysis)** 분석 기법 활용
- 이 논문에서는 합리적인 토픽의 개수를 결정하기 위하여, LDA 분석 기법 중 VEM(variational expectation-maximization) 알고리즘에 의해 토픽 수를 5개부터 30개까지 변화시키면서 클러스터링을 수행한 후 그 결과들에 대한 perplexity()함수 결과값을 도출하여 활용
- 또한 토픽의 해석 가능성, 의미 유용성 등을 고려하여 최종적으로 분석 대상의 토픽 수를 17개로 결정하여 LDA 분석을 실시하고 각 토픽을 대표하는 상위 단어들을 15개씩 추출

♣ 웹 스크래핑과 텍스트 마이닝 연구 수행 사례

4) 자료 분석



7

♣ 웹 스크래핑과 텍스트 마이닝 연구 수행 사례

5) 빈도수 상위 단어에 의한 Word Cloud 생성 결과

8

♣ R을 활용한 텍스트 마이닝 연구 수행 사례

6) 토픽 분석 결과

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
인터넷/ 모바일 콘텐츠	영상/음향 처리	정보 보호 및 보안	컴퓨터 네트워크	인터넷/ 모바일 콘텐츠
모니터링	영상	보안	네트워크	서비스
센서	검출	인증	노드	콘텐츠
네트워크	인식	서비스	무선	게임
서비스	알고리즘	서버	프로토콜	인터넷
선박	차원	rfid	센서	교육
스마트	얼굴	공격	라우팅	정보
정보	이미지	정보	에너지	학습
rfid	카메라	xml	패킷	스마트폰
설계	객체	파일	전송	분석
개발	물체	클라우드	mac	개발
통신	추적	네트워크	알고리즘	디지털
실시간	정보	프로토콜	경로	방송
스마트폰	이진화	암호화	클러스터	iptv
프로토콜	생성	스마트폰	트래픽	추천
led	홀로그램	설계	통신	학습자

♣ 웹 스크래핑과 텍스트 마이닝 연구 수행 사례

6) 토픽 분석 결과

Topic 6	Topic 7	Topic 8	Topic 9	Topic 10	Topic 11	Topic 12	Topic 13
센서 시스템	인공지능 및 지능 시스템	반도체(물성)	무선 통신	무선 통신	반도체(설계)	반도체(물성)	반도체(물성)
신호	알고리즘	회로	채널	주파수	하드웨어	안테나	박막
센서	제어기	전압	간섭	신호	설계	대역	mosfet
측정	퍼지	설계	네트워크	채널	암호	설계	문턱전압이하
검출	예측	전력	성능	ofdm	fpga	ghz	게이트
레이더	로봇	cmos	안테나	성능	복호	마이크로스트립	이중게이트
분석	신경회로망	변환기	전송	전송	연산	슬롯	비대칭
광섬유	프레임	주파수	mimo	변조	구조	광대역	문턱전압
개발	유전자	공정	기지국	반송파	프로세서	패치	변화
패턴	성능	출력	부호	비트	메모리	대역폭	전류
방사선	학습	위상	통신	위상	알고리즘	이득	스윙
gps	비선형	아날로그	용량	위상잡음	블록	구조	소자
도플러	분류	efuse	주파수	페이딩	비트	발진기	dgmosfet
차원	부호화	디지털	중계	동기	회로	주파수	산화막
자극	경로	전류	분석	링크	ldpc	기판	전압
음성	가중치	otp	노드	분석	알고리듬	소형	트랜지스터

♣ 웹 스크래핑과 텍스트 마이닝 연구 수행 사례

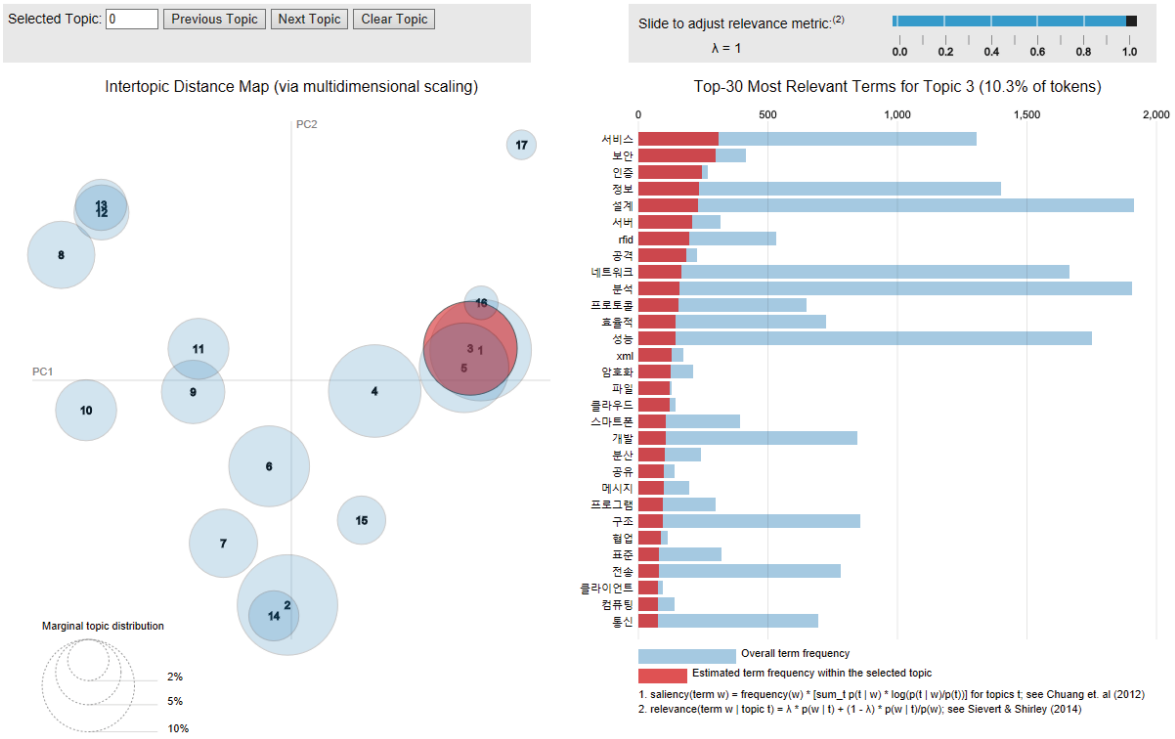
6) 토픽 분석 결과

Topic 14	Topic 15	Topic 16	Topic 17
영상/음향 처리	센서 신호 처리	지능형 차량 정보 처리	데이터 통신
잡음	위치	차량	voip
영상	태그	자동차	서비스
에지	rfid	운전자	스케줄링
알고리즘	실내	영상	for
임펄스	알고리즘	카메라	음성
검출	추정	지능형	network
마스크	정보	개발	qos
워터마킹	충돌	무선통신	mobile
가중치	오차	설계	품질
디지털	차량	무선	패킷
음성	gps	방지	시뮬레이터
제거	인식	스마트	실시간
열화	거리	진단	공격
변환	무선	usb	지연
방법들	슬롯	실시간	실행

11

♣ 웹 스크래핑과 텍스트 마이닝 연구 수행 사례

7) Intertopic Distance Map(IDM)



12

♣ 웹 스크래핑과 텍스트 마이닝 연구 수행 사례

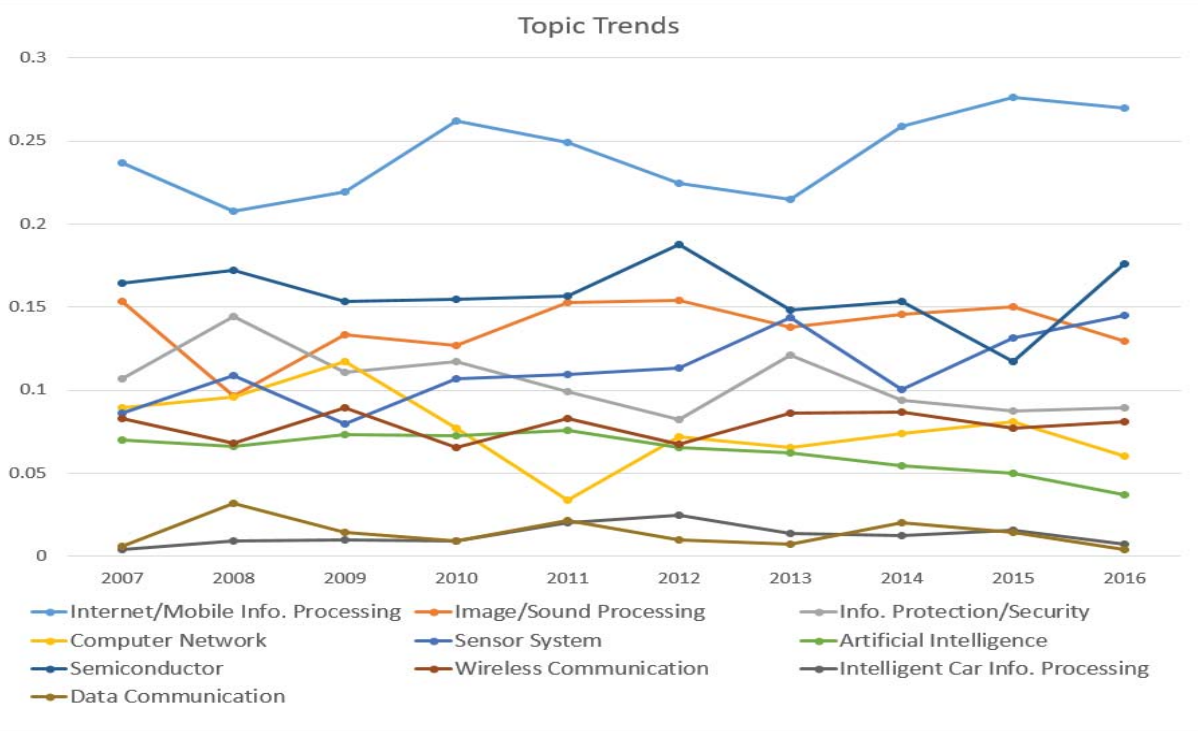
8) 최종 토픽 정리 결과

1. 인터넷/모바일 콘텐츠(Topic 1, 5)
2. 영상/음향 처리(Topic 2, 14)
3. 정보 보호 및 보안(Topic 3)
4. 컴퓨터 네트워크(Topic 4)
5. 센서 시스템(센서 신호 처리(Topic 6) + 센서 정보 처리(Topic 15))
6. 인공지능 및 지능 시스템(Topic 7)
7. 반도체(반도체(물성)(Topic 8, 12, 13) + 반도체(설계)(Topic 11))
8. 무선 통신(Topic 9, 10)
9. 지능형 차량 정보 처리(Topic 16)
10. 데이터 통신(Topic 17)

13

♣ 웹 스크래핑과 텍스트 마이닝 연구 수행 사례

9) Topic Trends



14

[1] LDA를 활용한 토픽 모델링

- **학습 목표**

1. 분석 데이터 수집을 위한 웹 데이터 스크래핑
2. 토픽 모델링을 위한 R 라이브러리의 종류 및 설치
3. 토픽 모델링을 위한 데이터 전처리 과정
4. 전체 단어 빈도수 분석 및 워드 클라우드 생성

1. 분석 데이터 수집을 위한 웹 데이터 스크래핑

- 분석을 위한 의료 관련 빅 데이터를 확보하기 위해서는 일반적으로 웹 문서를 활용함
- 컴퓨터 프로그램을 이용하여 웹 문서를 자동으로 수집하는 것을 "웹 스프래핑" 또는 "웹 크롤링"이라고 함

- 일반적으로 python 프로그램을 이용하여 웹 스크래핑을 수행
- 여기에서는 중앙일보 뉴스 사이트(https://news.joins.com/) 를 대상으로 '건강', '비만' 이라는 두 개의 키워드를 이용하여 나오는 기사들 중 1000개를 수집하여 실습에 활용

1. 분석 데이터 수집을 위한 웹 데이터 스크래핑

- 첫번째 프로그램: joongang_search_list2.py

중앙일보 검색 사이트에서 뉴스만 검색하여 제목과 사이트를 찾아서
아래와 같이 검색 결과를 주소와 제목을 쌍의 파일로 저장함

출력 파일 형태:

https://news.joins.com/article/22020746, "살 빼면 돈이 얼마나 ..."

https://news.joins.com/article/22021915, "우유 매일 12컵 ..."

https://news.joins.com/article/22022284, "건강한 당신 술 ..."

.....

1. 분석 데이터 수집을 위한 웹 데이터 스크래핑

- 두번째 프로그램: joongang_article_scraping2.py

기사 사이트 주소와 제목쌍으로 이루어진 csv 형식의 리스트 파일로부터
각 기사의 날짜, 제목, 본문을 csv 형식의 파일로 저장하는 프로그램

입력 파일 형태:
https://news.joins.com/article/22020746, "살 빼면 돈이 얼마나 절약될까"
https://news.joins.com/article/22021915, "우유 매일 12컵 꾸준히 마시면 ..."
.....

출력 파일 형태:
"adate", "atitle", "article"
"2017.10.17", " 살 빼면 돈이 얼마나 절약될까 ", " 사진 픽사베이 어떤 ... "
"2017.10.17", " 우유 매일 12컵 꾸준히 마시면 복부비만 ...", " 매일 ...”
.....

1. 분석 데이터 수집을 위한 웹 데이터 스크래핑

- 분석 데이터 파일의 기본 형식은 아래와 같은 csv 형식

"adate", "atitle", "article"
"2017.10.17", "살 빼면 돈이 얼마나 절약될까 ", "사진 픽사베이 어떤 ... "
"2017.10.17", "우유 매일 12컵 꾸준히 마시면 복부비만 ...", "매일 ..."
"2017.10.18", "건강한 당신 술 안 먹는데 지방간 ...", "건강 비타민 지난해..."
"2017.10.18", "식습관 관리해 체력 키우고 ...", "암 환자가 치료 후 수월하게 ..."
.....
.....
.....

- 첫 줄에는 각 필드를 의미를 나타내는 제목(헤더)이 존재하며, 그 다음 줄부터 해당 필드와 동일 개수의 데이터들이 **CSV 형식**으로 한 줄씩 나타남

2. 토픽 모델링을 위한 R 라이브러리의 종류 및 설치

```
#RStudio 실행한 후 필요한 기능이 있는 라이브러리를 불러 들임
#해당 기능을 사용하기 직전에 불러 들여도 됨
library(rJava)
library(KoNLP)
library(NLP)
library(stringr)
library(SnowballC)
library(servr)
library(tm)

#한글 형태소 분석에 사용할 사전 선택
#useSejongDic() #세종사전 선택 함수
useNIADic() #정보화진흥원(NIA) 사전 선택함수
```

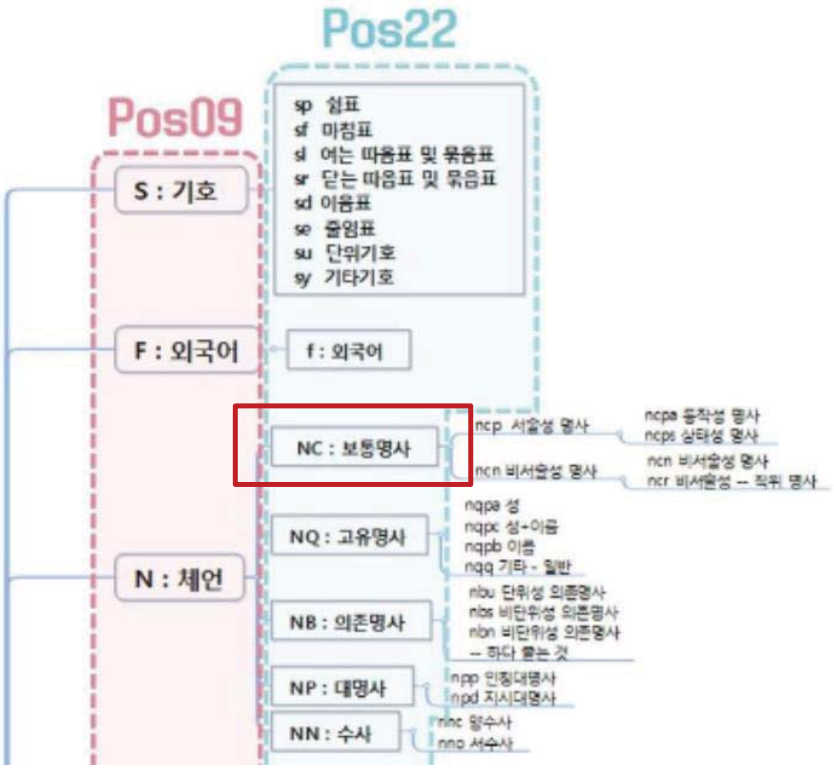
3. 토픽 모델링을 위한 데이터 전처리 과정

1) 한글 형태소 분석

```
while(cnt<=nws_len){ # 보통 명사만 추출하여 원 문장에 바꿔 넣음
    phr <- SimplePos22(nws$article[cnt])  # 22가지 품사를 추출하는 함수
    v1 <- paste(phr)
    d1 <- str_match(v1, '([가-힣]+)/NC')
    # 모든 한글에 대해서 보통 명사만 추출
    k1 <- d1[,2]
    av <- k1[!is.na(k1)]                    #보통 명사가 아닌 것을 삭제
    len=length(av)
    count=1
    conc <- ""
```


3. 토픽 모델링을 위한 데이터 전처리 과정

1) 한글 형태소 분석



4. 전체 단어 빈도수 분석 및 워드 클라우드 생성

- 빈도수 상위 단어 리스트 일부 예시

X	"freq"
말	1892
이상	1725
환자	1720
때	1610
건강	1609
교수	1601
사람	1525
경우	1310
질환	1244
운동	1237
치료	1127
몸	1081
위험	1067
결과	1039
음식	1023
체중	1004
비	946
비만	926
섭취	911
연구	877
효과	876
여성	864

23

4. 전체 단어 빈도수 분석 및 워드 클라우드 생성

- 워드 클라우드 생성 예 #1

24

4. 전체 단어 빈도수 분석 및 워드 클라우드 생성

- 워드 클라우드 생성 예 #2

[2] 토픽 추출 및 토픽 트랜드 분석

- **학습 목표**

1. 적절한 토픽 수 선정을 위한 전처리

2. 결정된 토픽 수에 의한 LDA 기반의 토픽 추출 및 분석
   1) 토픽별 상위 단어 추출 및 전체 비율 계산
   2) 추출된 토픽 결과 및 의미 해석
   3) 각 토픽의 기간별 비율 변화 추이
   4) ggplot 함수에 의한 기간별 토픽 비율 변화 가시화

1. 적절한 토픽수 선정을 위한 전처리

- 대량의 문서 데이터로부터 토픽 모델링을 수행하는 과정에서 처음으로 결정해야 하는 중요한 한 가지가 사전에 적절한 토픽수를 결정하는 것임(K-means 기법에서 K를 결정하는 것과 유사)
- 일반적으로 토픽수의 지정은 주로 산출되는 토픽들의 해석 가능성, 타당도, 연구 문제에 비추어 본 유용성 등에 의존하여 결정함
- 하지만 실제 분석을 위한 토픽수를 결정하는 것은 매우 어려운 문제로 많은 연구자들이 고민하고 있음
- R에서는 문서들의 군집화된 결과를 이용하여 군집화의 적절성을 평가해 주는 perplexity 함수를 제공함
- 여기에서는 perplexity 함수를 이용한 토픽 수를 결정하는 방법을 사용함

1. 적절한 토픽수 선정을 위한 전처리

1) perplexity 함수

- perplexity 함수는 군집내 밀집도와 군집간 거리 개념을 이용하여 혼잡도라는 개념의 수치를 산출
- perplexity 함수 결과값이 낮을수록 좋은 모델
- 이론적으로 이 함수 값이 가장 작아지는 수를 토픽수로 결정해야 하지만, 실제적으로는 문서들 간의 상관도가 대체로 높지 않아 매우 큰 토픽수에서 최소값을 가져 분석을 어렵게 함
- 따라서 일반적으로는 최소값은 아니더라도 함수 값의 추이를 보고 값의 변화가 적은 부분에서 토픽수를 결정하여 활용

1. 적절한 토픽수 선정을 위한 전처리

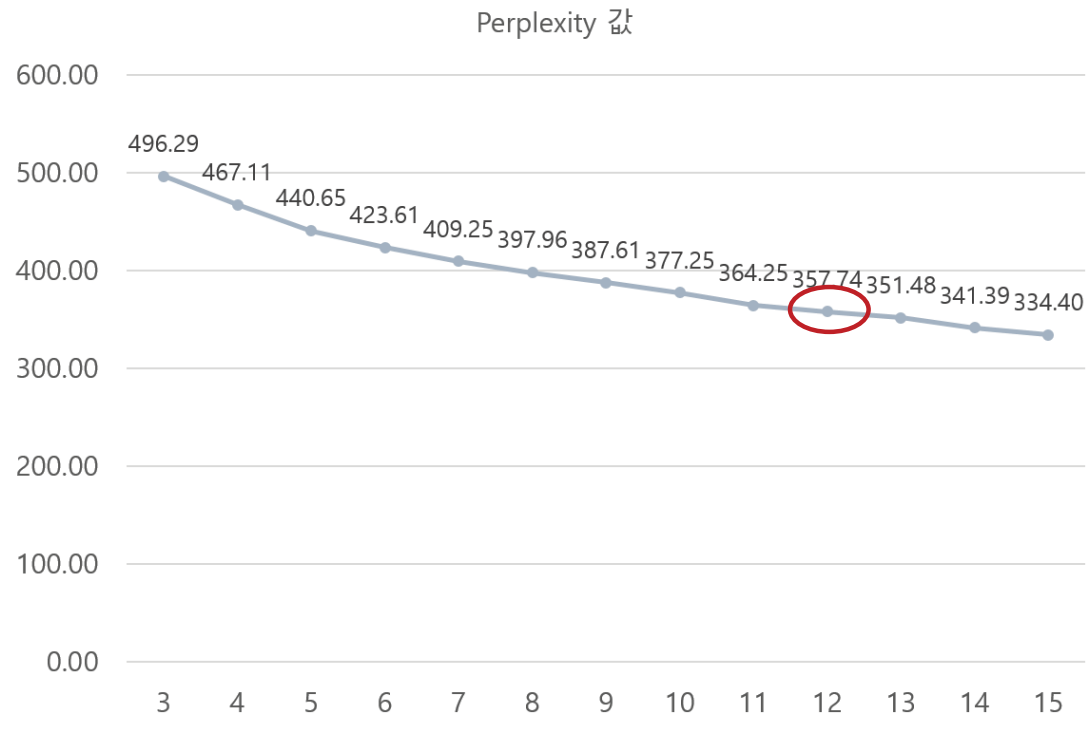
2) perplexity 함수 수행 결과 예(실습 샘플 데이터 활용)

토픽수	Perplexity 값
3	496.29
4	467.11
5	440.65
6	423.61
7	409.25
8	397.96
9	387.61
10	377.25
11	364.25
12	357.74
13	351.48
14	341.39
15	334.40

29

1. 적절한 토픽수 선정을 위한 전처리

3) perplexity 함수 수행 결과 예(실습 샘플 데이터 활용)



30

2. 결정된 토픽수에 의한 LDA 기반의 토픽 추출 및 분석

1) 추출된 토픽 결과 및 의미 해석

- Result_Topics(obesity_health)(500)12(20).txt

토픽번호	빈도수 상위 20개 단어
1	수술,환자,당뇨,위,치료,당뇨병,인슐린,소장,혈당,우회술,음식물,건강보험,췌장,위암,고도비,분비,상부,시스템,적용,형
2	트럼프,대통령,백악관,오바마,미셸,미국,햄버거,메뉴,당시,몸무게,무릎,콜레스테롤,권고,지난달,평가,고기,이용한,이날,북한,아동
3	관절,건선,통증,무릎,피부,관절염,연골,치료,통풍,염증,요산,천골,골관절염,도수치료,뼈,악화,퇴행성,전신,면역,완화
4	콜레스테롤,단백질,보이차,추출물,우유,섭취,효과,효소,지방,연구,제품,도움,영양소,성분,수치,그룹,건강,혈관,체내,감소
5	정부,한국,먹방,사진,국내,말,기자,올해,곱창,회,계획,건강,사업,시장,관련,국가,미국,생각,소비,규제

31

2. 결정된 토픽수에 의한 LDA 기반의 토픽 추출 및 분석

1) 추출된 토픽 결과 및 의미 해석 - 계속

토픽번호	빈도수 상위 20개 단어
6	성장,아이,성조숙증,때,키,부모,성장호르몬,비만,분비,초경,어린이,황기,자녀,자궁내막암,사춘기,소아청소년,근시,셀룰라이트,여아,엄마
7	대장암,다시보기,가공육,기내식,니켈,대장,발암물질,적색육,위암,날개,한국인,용종,검진,발생률,중앙일보,변비,사망률,간염,식이섬유,술
8	섭취,식품,과일,음식,설탕,당,쌀,영양,식단,식사,식습관,탄수화물,습식사료,나트륨,양,강아지,섭취량,수분,함량,영양소
9	지방간,알코올성,간암,지방간이,체험단,이소성,간염,허리둘레,운동량,동반,알코올,복부,형,조직,당분,회복,섭취하,외식,권장,요산
10	김치,유산균,아토피,장내,균,미생물,요리,김장,배추,소금,피부염,건강한,농산물,유해균,맛,약간,세균,대변,껍질,성분
11	이상,비,비만,남성,여성,대사증후군,비율,말,결과,경우,위험,지난해,성인,증가,영향,기자,연구팀,조사,비만율,교수
12	환자,때,운동,교수,사람,말,질환,경우,이상,몸,위험,체중,건강,증상,음식,정도,치료,발생,상태,후

32

2. 결정된 토픽수에 의한 LDA 기반의 토픽 추출 및 분석

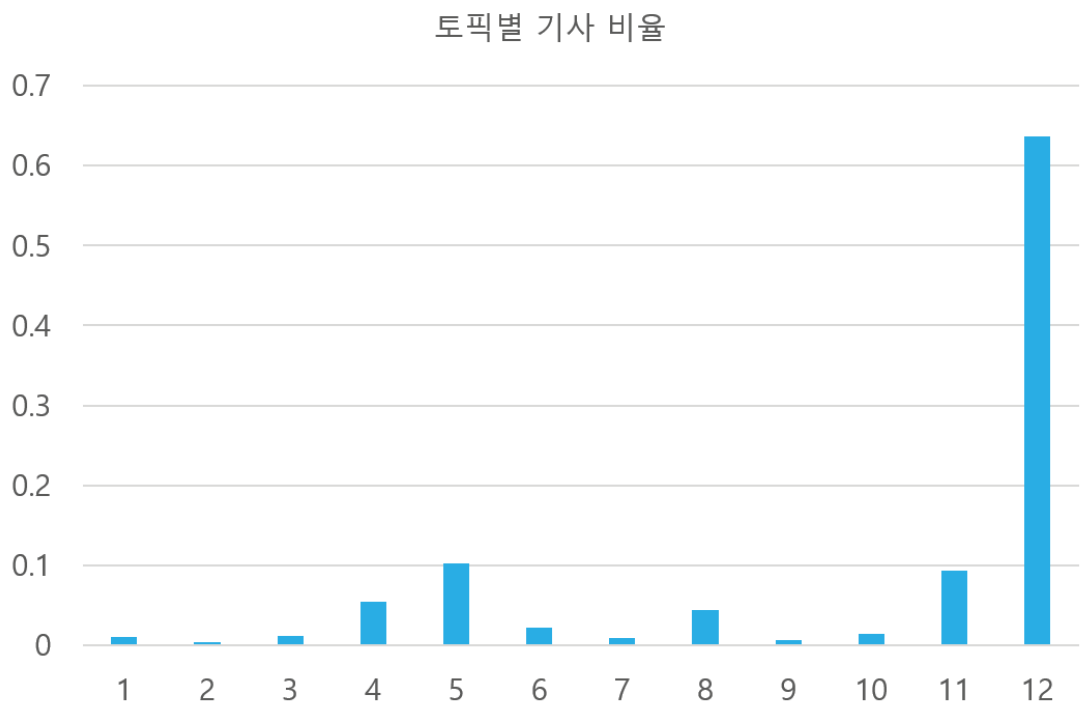
2) 추출된 토픽 결과 및 의미 해석

- Topic_proportion(obesity_health)(500)12(20).txt

토픽번호	토픽별 기사 비율
1	0.011
2	0.004
3	0.012
4	0.055
5	0.103
6	0.022
7	0.009
8	0.044
9	0.006
10	0.014
11	0.094
12	0.636

2. 결정된 토픽수에 의한 LDA 기반의 토픽 추출 및 분석

2) 추출된 토픽 결과 및 의미 해석



2. 결정된 토픽수에 의한 LDA 기반의 토픽 추출 및 분석

3) 각 토픽의 기간별 비율 변화 추이

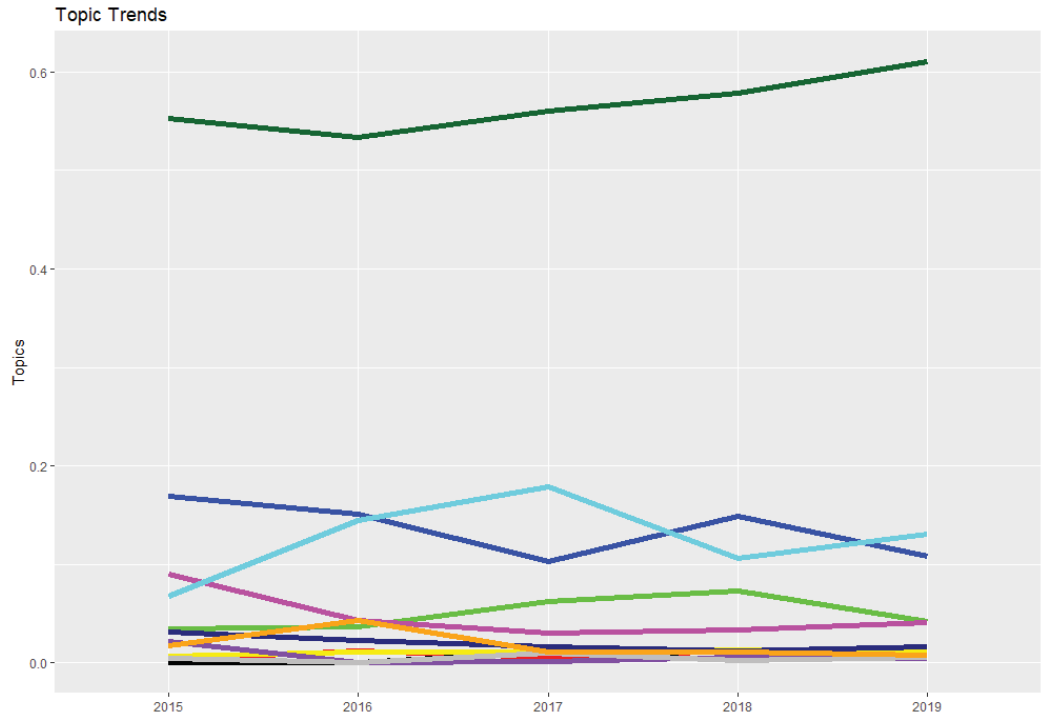
- Result_Topics(obesity_health)(500)12(20).txt

토픽번호/년도	1	2	3	4	5	6	7	8	9	10	11	12
2015	0.004	0.000	0.006	0.034	0.169	0.031	0.021	0.090	0.005	0.018	0.068	0.553
2016	0.012	0.000	0.012	0.037	0.151	0.023	0.000	0.043	0.000	0.043	0.144	0.534
2017	0.006	0.011	0.011	0.062	0.103	0.017	0.001	0.030	0.010	0.011	0.179	0.560
2018	0.012	0.003	0.014	0.073	0.149	0.013	0.006	0.033	0.002	0.011	0.107	0.578
2019	0.014	0.007	0.011	0.042	0.108	0.016	0.005	0.041	0.006	0.007	0.131	0.611

35

2. 결정된 토픽수에 의한 LDA 기반의 토픽 추출 및 분석

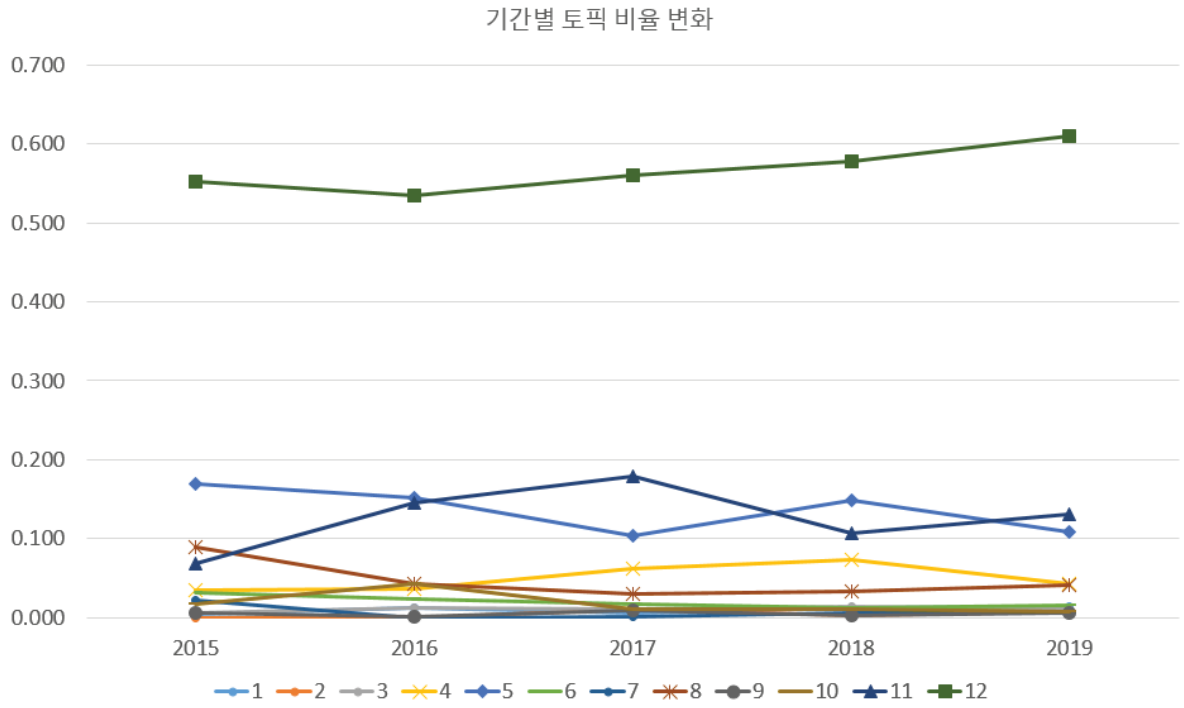
4) ggplot 함수에 의한 기간별 토픽 비율 변화 가시화



36

2. 결정된 토픽수에 의한 LDA 기반의 토픽 추출 및 분석

5) 각 토픽의 기간별 비율 변화 추이 그래프



[3] IDM 생성 및 분석

- **학습 목표**

1. IDM(Inter Distance Map) 개요
2. IDM 생성
3. IDM 분석 및 활용

1. IDM(Inter Distance Map) 개요

- 추출된 각 토픽들 간의 상관 관계를 2개의 차원(PC1, PC2)으로 축약하여 보여주는 시각화 그래프

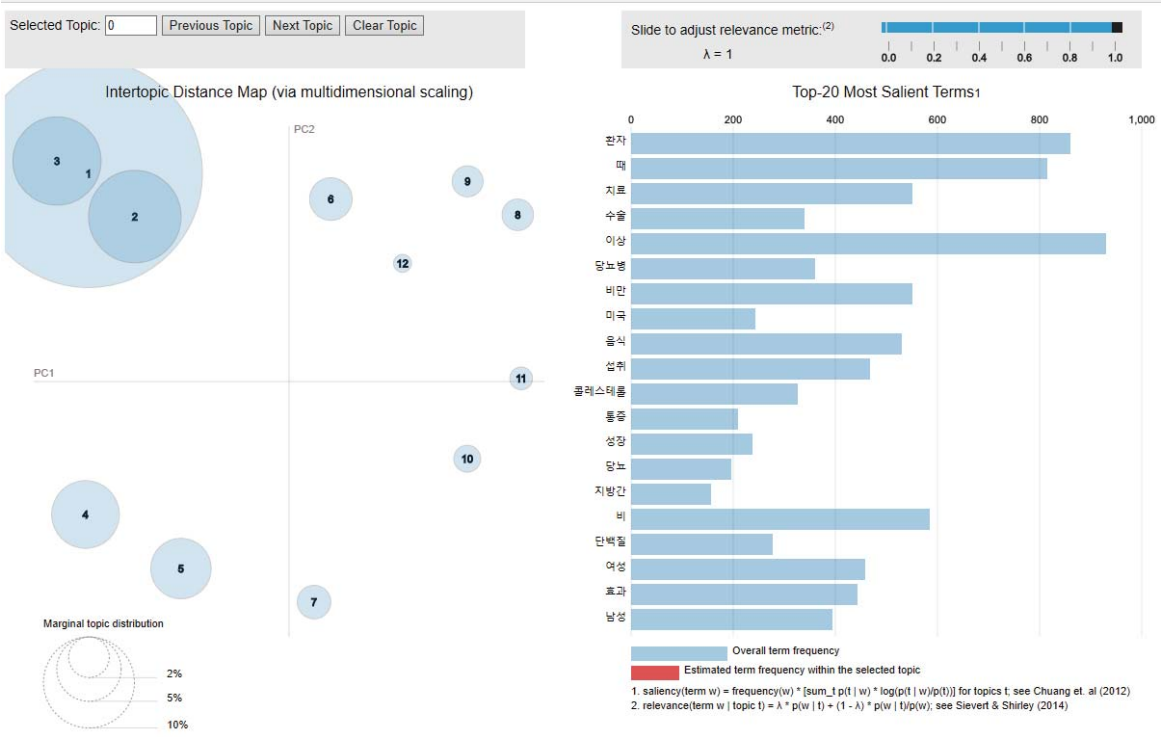
- 전체 문서 내에서 빈도수 상위 단어를 확인할 수 있으며, 또한 토픽별 상위 단어와 그 비율도 확인할 수 있는 그래프

- 각 토픽을 의미하는 원은 전체 문서에서의 비율이 면적에 비례하여 표출됨

- 원과 각 단어들에 대한 인터렉티브 기능을 가지고 있어 문서와 단어 간의 상관 관계, 비율을 확인할 수 있음

1. IDM(Inter Distance Map) 개요

- 실습 데이터로 생성한 IDM



2. IDM 생성

- IDM 생성 생성을 위해서는 dtm과 Gibbs sampling 방식에 의한 군집화 결과가 먼저 산출되어 있어야 함

- 여기서는 다시 한번 이 과정을 포함해서 전체 과정을 제시

- 생성된 결과는 웹 브라우저에 따라 표출 결과 여부나 형태가 달라질 수 있으므로 가급적 Microsoft Edge 브라우저를 이용

3. IDM 분석 및 활용

- 생성된 IDM은 추출된 각 토픽들 간의 상관 관계를 2개의 차원(PC1, PC2)으로 축약하여 보여주므로, perplexity에 의해 결정한 토픽 수를 다시 한번 검증해 보고 유사한 토픽을 하나로 합병할 수 있는 근거를 제시해 줌

- 앞에서 토픽 수를 결정하여 분석을 모두 실시하였다고 해도, IDM 형태에 따라 토픽을 합병함으로써 더욱 명확하고 설득력 있는 토픽 분석으로 업그레이드 할 수 있음

감사합니다.

Q/A

의사결정트리를 활용한 분류 기법

15:00~16:50

김희철 교수(인제대학교)

의사결정트리를 활용한 분류 기법

2019. 08. 23.

김희철

인제대학교 컴퓨터공학부/디지털항노화헬스케어학과

의사결정트리

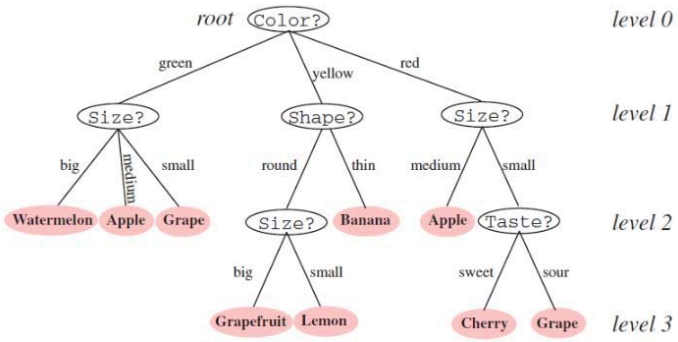
퍼지의사결정트리

의사결정트리 (Decision Tree)

- 정의

  의사결정규칙(decision rule)을 트리구조로 도표화하여 관심 대상이 되는 집단을
  몇 개의 소집단으로 분류(classification)하거나 예측(prediction)을 수행하는 분석방법

  스무고개와 개념이 유사 : 일련의 질문에 근거해서 데이터를 분류해주는 알고리즘



3

의사결정트리 (Decision Tree)

- 기계학습에서 대표적인 Supervised Learning 방법
- 목표 변수가 범주형인 경우 분류 트리, 수치형인 경우 회귀 트리
- 의사결정트리를 활용하는 분야

- 분류(Classification)
- 예측(Prediction)
- 세분화(Segmentation)
- 차원 축소, 변수선택, 연속 변수 이산화

- 응용 영역

- 시장조사, 광고조사, 의학연구, 품질관리 등의 다양한 분야에서 활용
- 특히 마케팅의 경우, 고객 타겟팅, 고객들의 신용점수화, 캠페인 반응분석,
  고객행동예측, 고객 세분화 등에 사용

4

의사결정트리 (Decision Tree)

▪ 장점
- Human readable한 트리구조로 표현되므로 결과를 쉽게 이해할 수 있음 (if..then.. 방식으로 결과 표현)
- 그 구조가 단순하여 해석이 용이

▪ 단점
- 분류 기준값의 경계선 근방의 자료 값에 대해서는 오차가 클 수 있으며(비연속성),
- 로지스틱 회귀와 같이 각 예측변수의 효과를 파악 어려움
- Hill climbing 방식 및 Greedy 방식을 사용하는데, 일반적으로 Greedy 방식의 알고리즘은 local optimization에 빠지기 쉬움

5

학습 데이터 사례

Day	Outlook	온도	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

6

의사결정트리 알고리즘

- 대표적 알고리즘

- CHAID(Kass, 1980)
- CART(Breiman 등, 1984)
- ID3(Quinlan, 1986)
- C4.5(Quinlan, 1993)
- C5.0(Quinlan, 1998)
- 이들의 장점을 결합한 다양한 알고리즘

ID3 알고리즘

- 대표적인 의사결정트리 기반 분류 알고리즘

1. 전체 데이터를 포함하는 루트 노드를 생성한다.

2. 만약 샘플들이 모두 같은 클래스라면, 노드는 Leaf 노드가 되고,
   해당 클래스로 레이블을 부여한다. (종료 조건)

3. 그렇지 않으면 정보이득(Information Gain)이 높은
   (즉, 데이터를 가장 잘 구분할 수 있는) 속성을 선택한다.
   (이때 정보이득은 엔트로피(Entropy)의 변화를 가지고 계산한다.)

4. 선택된 속성으로 가지(Branch)를 뻗어 하위 노드들을 생성한다.
   (각 하위 노드들은 가지의 조건을 만족하는 레코드들이다.)

5. 각 노드에 대하여 2단계로 이동한다.

ID3 알고리즘 (예제를 통한 이해)

- 대표적인 의사결정트리 기반 분류 알고리즘
- 학습 예제를 가장 잘 분류할 수 있는 속성을 루트에 둠
  · 엔트로피(Entropy), 정보 이득(Information Gain)
- 엔트로피 : 주어진 데이터 집합의 혼잡도를 의미
  Entropy(S) = − P(+) * $\log_2$ (P(+)) − P(−) * $\log_2$ (P(−))
  P(+) : 긍정 레코드의 개수 / 총 레코드의 개수
  P(−) : 부정 레코드의 개수 / 총 레코드의 개수

  Entropy(S) 구하기

  $S = PlayTennis[9+, 5-]$
  $P(+) = \frac{9}{14}$
  $P(-) = \frac{5}{14}$
  $Entropy(S) = -\frac{9}{14} \times log_2 \frac{9}{14} - \frac{5}{14} \times log_2 \frac{5}{14} = 0.940$

ID3 알고리즘

- 정보 이득 : 데이터를 분할하기 전과 후의 변화
  앞서 구한 Label Entropy인 Entropy(S)와
  Node의 Entropy의 차이를 구함

$$Gain(S, A) \equiv Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

1. Gain(S, Outlook) 구하기

$Values(Outlook) = Sunny, Overcast, Rain$
$S_{Sunny} = [2+, 3-]$, $S_{Overcast} = [4+, 0-]$, $S_{Rain} = [3+, 2-]$

Outlook 노드에서 Rain이라는 값을 가진 레코드는 총 5개이고 5개 중에서 PlayTennis가 Yes인 경우가 3개, No인 경우가 2개

$Gain(S, Outlook) = Entropy(S) - \sum_{v \in Sunny, Overcast, Rain} \frac{Sv}{S} Entropy(Sv)$
$= Entropy(S) - \frac{5}{14} \times Entropy(S_{Sunny}) - \frac{4}{14} \times Entropy(S_{Overcast}) - \frac{5}{14} \times Entropy(S_{Rain})$
$= 0.940 - \frac{5}{14} \times 0.971 - \frac{4}{14} \times 0 - \frac{5}{14} \times 0.971$
$= 0.246$

ID3 알고리즘

2. Gain(S, Wind) 구하기

$Values(Wind) = Weak, Strong$
$S_{Weak} = [6+, 2-]$
$S_{Strong} = [3+, 3-]$
$Gain(S, Wind) = Entropy(S) - \sum_{v \in \{Weak, Strong\}} \frac{Sv}{S} Entropy(Sv)$
$= Entropy(S) - \frac{8}{14} \times Entropy(S_{Weak}) - \frac{6}{14} \times Entropy(S_{Strong})$
$= 0.940 - \frac{8}{14} \times 0.811 - \frac{6}{14} \times 1.00$
$= 0.048$

11

ID3 알고리즘

3. Gain(S, Humidity) 구하기

$Values(Humidity) = High, Normal$
$S_{High} = [3+, 4-]$
$S_{Normal} = [6+, 1-]$
$Gain(S, Humidity) = Entropy(S) - \sum_{v \in \{High, Normal\}} \frac{Sv}{S} Entropy(Sv)$
$= Entropy(S) - \frac{7}{14} \times Entropy(S_{High}) - \frac{7}{14} \times Entropy(S_{Normal})$
$= 0.940 - \frac{7}{14} \times 0.985 - \frac{7}{14} \times 0.592$
$= 0.151$

12

ID3 알고리즘

Outlook 노드의 Information Gain : 0.246
Wind 노드의 Information Gain : 0.048
Humidity 노드의 Information Gain : 0.151

따라서 Information Gain이 가장 높은
"Outlook" 노드를 Root Node로 선택

13

완성된 트리 사례

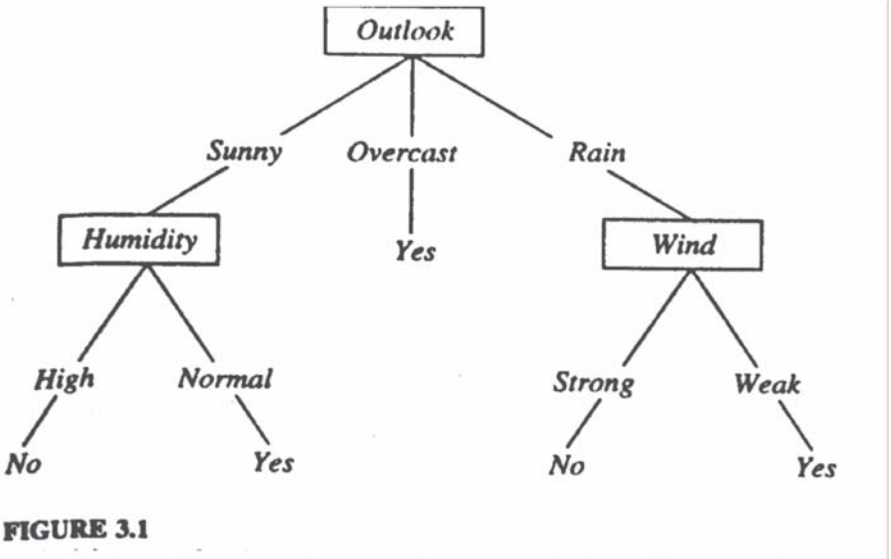


FIGURE 3.1

14

예제 프로그램

화장품 구매여부 예측하기

	A	B	C	D	E	F	G
1	고객번호	성별	나이	직장여부	결혼여부	차량보유여부	구매여부
2	1	남	30대	NO	YES	NO	NO
3	2	여	20대	YES	YES	YES	NO
4	3	여	20대	YES	YES	NO	NO
5	4	여	40대	NO	NO	NO	NO
6	5	여	30대	NO	YES	NO	NO
7	6	여	30대	NO	NO	YES	NO
8	7	여	20대	NO	YES	NO	NO
9	8	여	20대	NO	YES	YES	YES
10	9	여	30대	YES	YES	NO	YES
11	10	남	40대	YES	NO	YES	NO
12	11	남	20대	NO	NO	NO	NO
13	12	남	30대	NO	YES	YES	NO
14	13	남	20대	YES	NO	NO	NO
15	14	여	30대	YES	YES	NO	YES
16	15	남	30대	YES	YES	YES	YES
17	16	여	30대	YES	NO	NO	NO
18	17	여	30대	NO	YES	YES	YES
19	18	남	20대	YES	YES	NO	NO
20	19	남	40대	YES	NO	YES	NO

15

예제 프로그램

고객번호	성별	나이	직장여부	결혼여부	차량보유여부
31	남	30대	NO	YES	NO

Test Data

정보 이득이 가장 높은 "결혼여부" 노드를 루트로 선택

고객번호	성별	나이	직장여부	결혼여부	차량보유여부	구매여부
1	남	30대	NO	YES	NO	NO
2	여	20대	YES	YES	YES	NO
3	여	20대	YES	YES	NO	NO
4	여	40대	NO	NO	NO	NO
5	여	30대	NO	YES	NO	NO
6	여	30대	NO	NO	YES	NO
7	여	20대	NO	YES	NO	NO
8	여	20대	NO	YES	YES	YES
9	여	30대	YES	YES	NO	YES
10	남	40대	YES	NO	YES	NO
11	남	20대	NO	NO	NO	NO
12	남	30대	NO	YES	YES	NO
13	남	20대	YES	NO	NO	NO
14	여	30대	YES	YES	NO	YES
15	남	30대	YES	YES	YES	YES
16	여	30대	YES	NO	NO	NO
17	여	30대	NO	YES	YES	YES
18	남	20대	YES	YES	NO	NO
19	남	40대	YES	NO	YES	NO
20	여	40대	YES	YES	NO	YES
21	여	20대	NO	YES	YES	YES
22	남	30대	NO	NO	NO	NO
23	여	30대	YES	YES	NO	YES
24	남	30대	YES	NO	YES	NO
25	여	40대	NO	YES	YES	YES
26	남	30대	NO	YES	NO	NO
27	여	30대	YES	YES	YES	YES
28	여	40대	YES	NO	YES	NO
29	남	40대	YES	YES	NO	YES
30	여	40대	YES	YES	NO	YES

현재 Entropy : 0.9709505944546686
성별의 InformationGain : 0.11629598989239365
나이의 InformationGain : 0.02816996444958253
직장여부의 InformationGain : 0.019819440750390438
결혼여부의 InformationGain : 0.32365019815155627
차량보유여부의 InformationGain : 0.008690515487248751
선택된 노드 : 결혼여부
값 : YES

16

예제 프로그램

고객번호	성별	나이	직장여부	결혼여부	차량보유여부	구매여부
1	남	30대	NO	YES	NO	NO
2	여	20대	YES	YES	YES	NO
3	여	20대	YES	YES	NO	NO
5	여	30대	NO	YES	NO	NO
7	여	20대	NO	YES	NO	NO
8	여	20대	NO	YES	YES	YES
9	여	30대	YES	YES	NO	YES
12	남	30대	NO	YES	YES	NO
14	여	30대	YES	YES	NO	YES
15	남	30대	YES	YES	YES	YES
17	여	30대	NO	YES	YES	YES
18	남	20대	YES	YES	NO	NO
20	여	40대	YES	YES	NO	YES
21	여	20대	NO	YES	YES	YES
23	여	30대	YES	YES	NO	YES
25	여	40대	NO	YES	YES	YES
26	남	30대	NO	YES	NO	NO
27	여	30대	YES	YES	YES	YES
29	남	40대	YES	YES	NO	YES
30	여	40대	YES	YES	NO	YES

현재 Entropy : 0.9709505944546686
성별의 InformationGain : 0.0912774462416801
나이의 InformationGain : 0.209986547010974
직장여부의 InformationGain : 0.06002335238512835
결혼여부의 InformationGain : 0.0
차량보유여부의 InformationGain : 0.0464393446710154
선택된 노드 : 나이
값 : 30대

정보 이득이 가장 높은 "나이"를 다음 노드를 루트로 선택

Entropy와 Information Gain을 계산해가며 "구매 여부"라는 Label이 같을 때 까지 반복

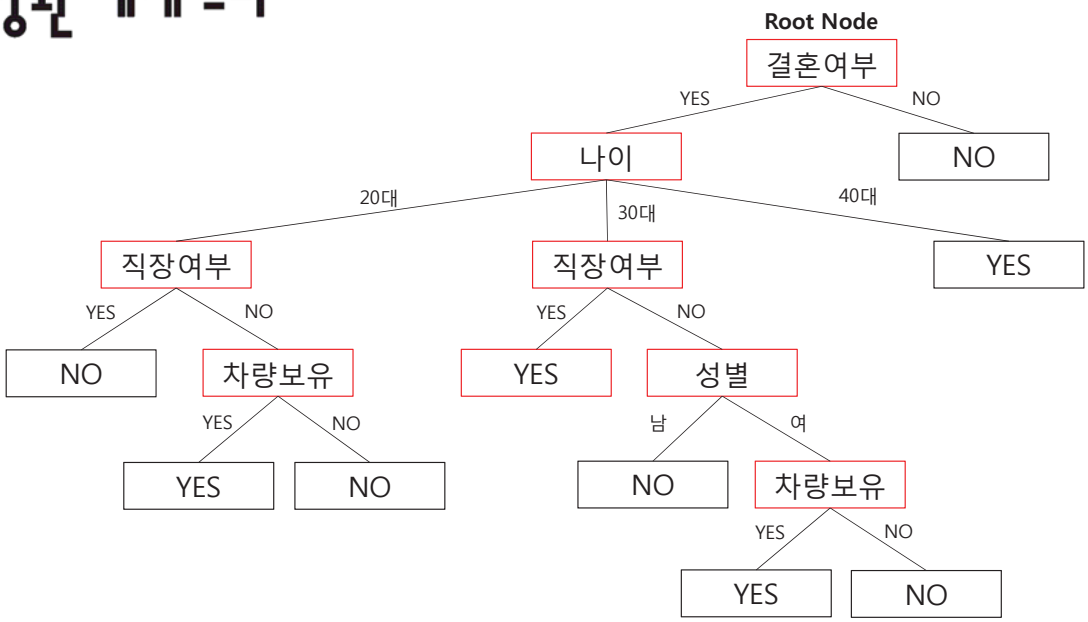
.
.
.

고객번호	성별	나이	직장여부	결혼여부	차량보유여부	구매여부
1	남	30대	NO	YES	NO	NO
12	남	30대	NO	YES	YES	NO
26	남	30대	NO	YES	NO	NO

분류 결과 : NO

최종적으로 "NO"라는 분류 결과를 통해
Test Data는 구매하지 않는다고 예측

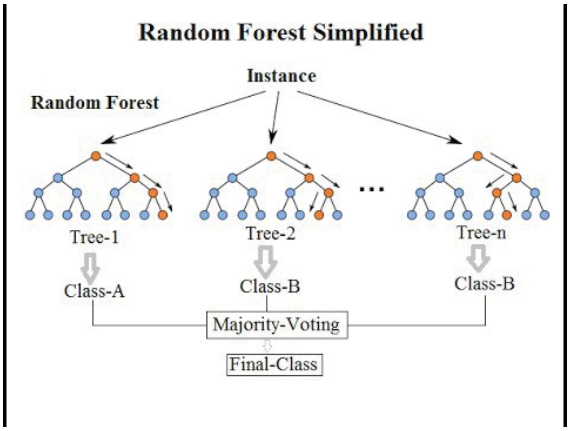
완성된 예제 트리



랜덤 포레스트 (Random Forest)

- 랜덤 포레스트란?

의사결정트리의 단점을 개선하기 위한 최적의 방법
- DT의 가장 심각한 문제는 Overfitting 문제로, 이 문제가 심각하면 일반화하여 사용하기 어려우며 랜덤 포레스트는 이를 해결하기 위한 방법으로 제안 됨 (2001년 Leo Breiman)
- 랜덤포레스트는 여러 개의 의사결정나무들을 생성한 다음, 각 개별 트리의 예측값들 중에서 가장 많은 선택을 받은 클래스(또는 label)로 예측하는 알고리즘
- 대표적 앙상블 학습 방법 중 하나

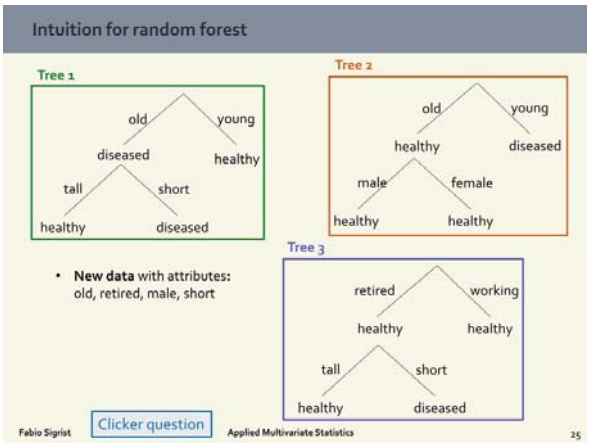


19

랜덤 포레스트 (Random Forest)

- 트리 생성 방법

- 랜덤 포레스트는 트리를 생성할 때, 각 노드는 랜덤하게 특성(feature)의 서브셋(자식 노드를)을 만들어 분할
- 중복을 허용하는 리샘플링(resampling) 즉, 부트스트래핑(bootstraping) 방식인 배깅(Bagging) 방식을 사용하기 때문에, 전체 학습 데이터셋에서 어떠한 데이터 샘플은 여러 번 샘플링 되고, 또 어떠한 샘플은 전혀 샘플링 되지 않을 수가 있음.
- 평균적으로 학습 단계에서 전체 학습 데이터셋 중 63% 정도만 샘플링 되며, 샘플링 되지 않은 나머지 37% 데이터 샘플을 **OOB(out-of-bag) 샘플**이라 함



20

랜덤 포레스트 (Random Forest)

- 트리생성 방법

  - 부트스트래핑 과정을 통해 N개의 샘플링 데이터 셋을 생성
  - 각 샘플링된 데이터 셋에서 임의의 특징을 선택하는 과정 진행
  - M개의 총 변수들 중에서 sqrt(M) 또는 M/3개의 개수만큼 변수들을 랜덤하게 선택하고 나머지 변수는 모두 제거하는 과정 반복
  - 특징선택이 진행된 의사결정트리들을 종합하여 앙상블 모델을 만들고 OOB 샘플을 검증셋(validation set)으로 사용

- 특징 중요도 결정

  - 특징(feature)의 상대적인 중요도를 측정하기 쉬움

- 의의

  - 각 특징선택의 임의성과 배깅을 통해 각 트리들의 예측들이 비상관화가 되어 일반화 성능 향상
  - 노이즈가 포함된 데이터에 이용하기 좋음
  - 데이터 셋 내의 데이터 분포가 고르지 않은 경우에도 사용됨

21

퍼지 의사결정트리과 비선형 분류기법

퍼지의 개념

- 일반 집합은 “어떤 조건들을 만족하는 것들의 모임” 으로 정의되며 경계가 확실히 구분되어 있다는 의미에서 크리스프(crisp) 집합이라고 한다.
- 일반 집합은 포함되는 경우에는 1, 포함되지 않는 경우에는 0으로 나타내는데 막연하거나 모호한 상태를 표현하기에 부적절한 상황들이 존재한다.

  ex) 키가 180이상인 남자들을 “키가 큰 사람들의 집합”이라고 했을 때, 집합에 포함되지 않는 키가 179인 남자는 키가 작다고 볼 수 있을까?
- 퍼지는 기존의 일반 집합이 표현할 수 없었던 불분명한 상태, 모호한 상태를 참 혹은 거짓의 이진 논리에서 벗어난 다치성으로 표현하는 논리 개념이다.

퍼지 집합

- 퍼지 집합의 명제는 참 또는 거짓이 아니라 부분적으로 참, 부분적으로 거짓이기 때문에 모호한 개념을 표현할 수 있다.
- 부분의 정도는 “소속도 " 라고 하여 보통 [0,1] 범위의 실수값으로 표현한다.
- “키가 큰 남자” 라는 퍼지 집합의 원소는 모든 남자지만, 이 집합의 소속도는 표에서 보듯 키에 좌우된다.
- 키가 205cm인 마크의 소속도는 1, 키가 152cm인 피터의 소속도는 0이고 그 사이에 있는 사람들은 1과 0사이의 소속도를 가지며, 부분적으로 크고 작다고 볼 수 있다.

이름	키(CM)	소속도	
		크리스프	퍼지
크리스	208	1	1.00
마크	205	1	1.00
존	198	1	0.98
톰	181	1	0.82
데이비드	179	0	0.78
마이크	172	0	0.24
밥	167	0	0.15
스티븐	158	0	0.06
빌	155	0	0.01
피터	152	0	0.00

퍼지 집합

- 퍼지 집합은 어떤 요소가 그 집합에 어느 정도 포함되는지 그 소속도를 나타내기 위해 소속 함수(Membership Function)을 사용한다.
- 소속 함수는 집합의 안쪽을 1, 바깥쪽을 0으로 하고 경계에서는 그 사이 값을 취하여 안쪽에 가까울수록 1에 가까운 값을 가진다.
- 퍼지 집합은 소속 함수 μ를 사용하여 다음과 같이 정의된다.

$$\mu_A(x) \rightarrow [0,1]$$

$\mu_A(x) = 1$	x가 완전히 A에 속한 경우
$\mu_A(x) = 0$	x가 완전히 A에 속하지 않은 경우
$0 < \mu_A(x) < 1$	x가 부분적으로 A에 속한 경우

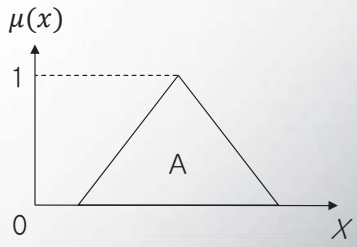
퍼지 집합 연산

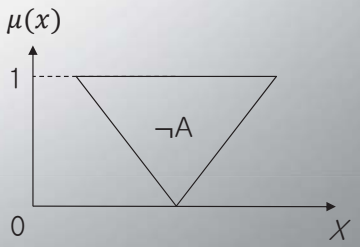
- 여집합(Complement)

크리스프 집합 : 어떤 원소가 그 집합에 속하지 않을까?

퍼지 집합 : 어떤 원소가 그 집합에 "얼마만큼" 속하지 않을까?

$$\boldsymbol{\mu_{\neg A}(x) = 1 - \mu_A(x)}$$





키가 큰 남자의 퍼지 집합 = (0/180, 0.25/182.5, 0.5/185, 0.75/187.5, 1/190)

키가 크지 않은 남자의 퍼지 집합 = (1/180, 0.75/182.5, 0.5/185, 0.25/187.5, 0/190)

퍼지 집합 연산

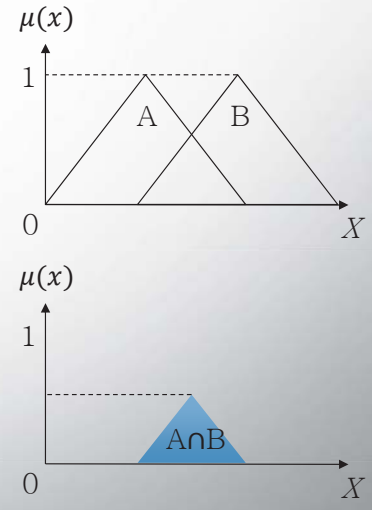
- 교집합(Intersection)

크리스프 집합 : 어느 원소가 두 집합에 모두 속할까?

퍼지 집합 : 원소가 두 집합 모두에 얼마만큼 속할까?

크리스프 집합의 교집합은 두 집합이 겹치는 영역을 말하지만, 퍼지 집합은 한 원소가 두 집합에 대해 서로 다른 정도로 속하기 때문에 각 원소의 퍼지 교집합에 대한 소속값은 두 집합에 대한 소속값 중 낮은 값이 된다.

$$\mu_{A\cap B}(x) = \min[\mu_A(x), \mu_B(x)] = \mu_A(x) \cap \mu_B(x)$$



키가 큰 남자의 퍼지 집합 = (0/165, 0/175, 0/180, 0.25/182.5, 0.5/185, 1/190)

키가 보통인 남자의 퍼지 집합 = (0/165, 1/175, 0.5/180, 0.25/182.5, 0/185, 0/190)

키가 큰 남자 ∩ 키가 보통인 남자 = (0/165, 0/175, 0/180, 0.25/182.5, 0/185, 0/190)

= (0/180, 0.25/182.5, 0/185)

퍼지 집합 연산

- 합집합(Union)

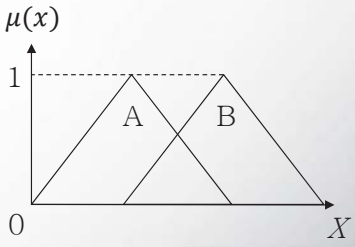
크리스프 집합 : 원소가 두 집합에 어느 쪽이든 속할까?

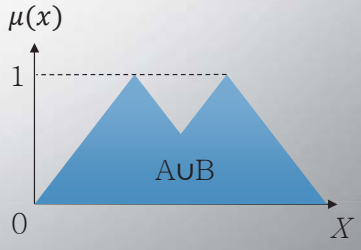
퍼지 집합 : 원소가 두 집합 어느 쪽이든 얼마만큼 속할까?

크리스프 집합의 합집합은 두 집합의 전체 원소를 포함하지만, 퍼지 집합의 합집합은 교집합의 반대이다.

즉, 각 원소의 합집합에 대한 소속값은 두 집합에 대한 소속값 중 높은 값이다.

$$\mu_{A\cup B}(x) = \max[\mu_A(x), \mu_B(x)] = \mu_A(x) \cup \mu_B(x)$$





키가 큰 남자의 퍼지 집합 = (0/165, 0/175, 0/180, 0.25/182.5, 0.5/185, 1/190)

키가 보통인 남자의 퍼지 집합 = (0/165, 1/175, 0.5/180, 0.25/182.5, 0/185, 0/190)

키가 큰 남자 ∪ 키가 보통인 남자 = (0/165, 1/175, 0.5/180, 0.25/182.5, 0.5/185, 1/190)

퍼지 의사결정트리

- 퍼지 의사결정트리는 기존의 의사결정트리에서 확장된 방법이며 불확실한 분류 문제에서 지식을 추출하는 효과적인 방법
- 퍼지 이론을 적용하여 데이터 집합을 표현하고 트리 구조를 결정
- 기존의 의사결정트리와 동일한 접근법을 사용

  : 리프 노드에 도달하거나 속성 또는 레코드가 남아 있지 않을 때 까지 반복
- 분할 과정에서 퍼지 엔트로피와 퍼지 데이터 세트의 정보 이득을 계산하여 트리를 확장하기 위해 트리의 테스트 노드에서 사용할 속성을 선택

퍼지 의사결정트리

- 기존의 의사결정트리의 특징은 각 레코드가 특정 노드에 대해서 확실하게 속해 있거나 속해 있지 않지만 퍼지의 경우에는 다름
- 각 속성에 대해 언어변수를 정의하고, 주어진 예제의 소속도를 결정해야 함

온도
x_1 = 8
x_2 = 10
x_3 = 15
x_4 = 20
x_5 = 23
x_6 = 25

➡

온도		
시원함	화창함	더움
0.7	0.3	0.0
0.5	0.5	0.0
0.0	1.0	0.0
0.0	0.8	0.2
0.0	0.5	0.5
0.0	0.0	1.0

예제

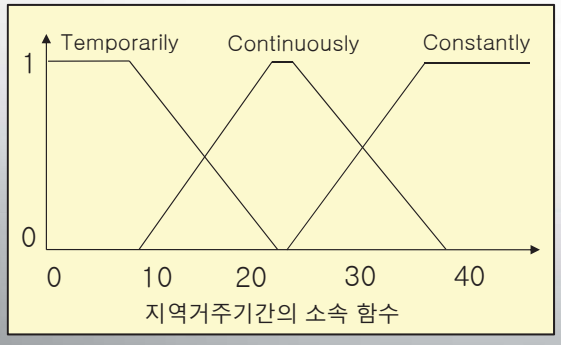
- 퍼지 의사결정트리 구축을 위한 트레이닝 셋

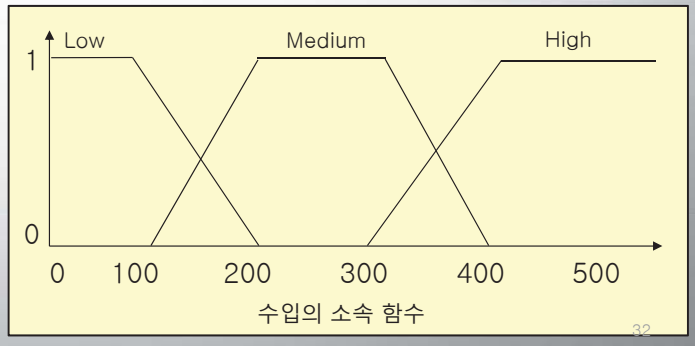
No.	지역거주기간(년)	수입(만원)	신용도
D1	1	100	0.0
D2	10	150	0.0
D3	15	200	0.1
D4	20	300	0.3
D5	30	250	0.7
D6	40	350	0.9
D7	40	500	1.0

- 지역 거주 기간이 25년이며 수입이 320만원인 고객이 은행으로 부터 대출을 받을 수 있을지에 대한 문제를 퍼지 의사결정 트리를 통해 분류

예제

- 속성에 대한 언어변수:
  - "지역거주기간" 은 *"temporarily", "continuously", "constantly";*
  - "수입" 은 *"Low", "medium", "high".*





트레이닝 셋의 퍼지화

- 소속 함수를 통해 변경된 속성들의 소속도

No.	지역거주기간			수입		
	Temporarily	Continuously	Constantly	Low	Medium	High
D1	1	0	0	1	0	0
D2	0.8	0.2	0	0.6	0.4	0
D3	0.5	0.5	0	0.1	0.9	0
D4	0.2	0.8	0	0	1	0
D5	0	0.5	0.5	0	1	0
D6	0	0	1	0	0.6	0.4
D7	0	0	1	0	0	1

퍼지 의사결정트리 구현

1. 총 엔트로피 구하기

- P_{yes} = 0 + 0 + 0.1 + 0.3 + 0.7 + 0.9 + 1.0 = 3 *("신용도" 속성의 positive examples)*
- P_{no} = 1 + 1 + 0.9 + 0.7 + 0.3 + 0.1 + 0 = 4 *("신용도" 속성의 negative examples, P_{yes}의 여집합)*
- $P = P_{yes} + P_{no}$ = 3 + 4 = 7 *(all examples)*

아래 식을 통해 전체 엔트로피 $E(S^N)$ 을 계산 :

$$E(S^N) = -\frac{3}{7} \cdot \log_2 \frac{3}{7} - \frac{4}{7} \cdot \log_2 \frac{4}{7} \approx 0,985 \text{ bit.}$$

퍼지 의사결정트리 구현

2. 지역거주기간 속성의 temporarily에 대한 엔트로피 구하기

E(S^N, 지역거주기간):

- $P_{yes}^{temporarily}$ = min (0, 1) + min (0; 0.8) + min (0.1; 0.5) + min (0.3; 0.2) + min (0.7, 0) + min (0,9; 0) + min (1, 0) = 0 + 0 + 0.1 + 0.2 + 0 + 0 + 0 = 0.3
- $P_{no}^{temporarily}$ = min (1, 1) + min (1; 0.8) + min (0.9; 0.5) + min (0.7; 0.2) + min (0.3, 0) + min (0,1; 0) + min (0, 0) = 1 + 0.8 + 0.5 + 0.2 + 0 + 0 + 0 = 2.5
- $P^{temporarily}$ = 0.3 + 2.5 = 2.8

따라서, E(지역거주기간, temporarily) $= -\frac{0,3}{2,8} \cdot \log_2 \frac{0,3}{2,8} - \frac{2,5}{2,8} \cdot \log_2 \frac{2,5}{2,8} \approx 0,491$ bit.

퍼지 의사결정트리 구현

- Temporarily와 마찬가지로 Continuosly, Constantly도 동일한 계산을 수행하면 아래 표와 같은 결과를 얻을 수 있음

	Temporarily	Continuously	Constantly
P_{yes}	0.3	0.9	2.4
P_{no}	2.5	1.7	0.4
E (bit)	0.491	0.931	0.592

퍼지 의사결정트리 구현

3. “지역거주기간” 속성의 총 엔트로피 계산:

- $E(S^N, \text{지역거주기간}) = \frac{2,5}{7} \cdot 0,491 + \frac{1,7}{7} \cdot 0,931 + \frac{0,4}{7} \cdot 0,592 \approx 0,486$ бит.

4. “지역거주기간” 속성에 대한 “정보 이득” 계산: <총 엔트로피 – 속성 엔트로피>

- $G(S^N, \text{지역거주기간})$ = 0.985 – 0.486 = 0.499 bit.

“수입” 속성에 대해서도 동일한 계산을 수행:

- $E(S^N, \text{수입})$ = 0.416 bit;
- $G(S^N, \text{수입})$ = 0.569 bit.

정보 이득 계산 결과 “수입“ 속성의 정보 이득이 가장 높으므로 “수입” 속성을 트리의 루트 노드로 선택

5. 수입의 속성에 대한 각 지역거주기간 속성들의 소속도를 계산

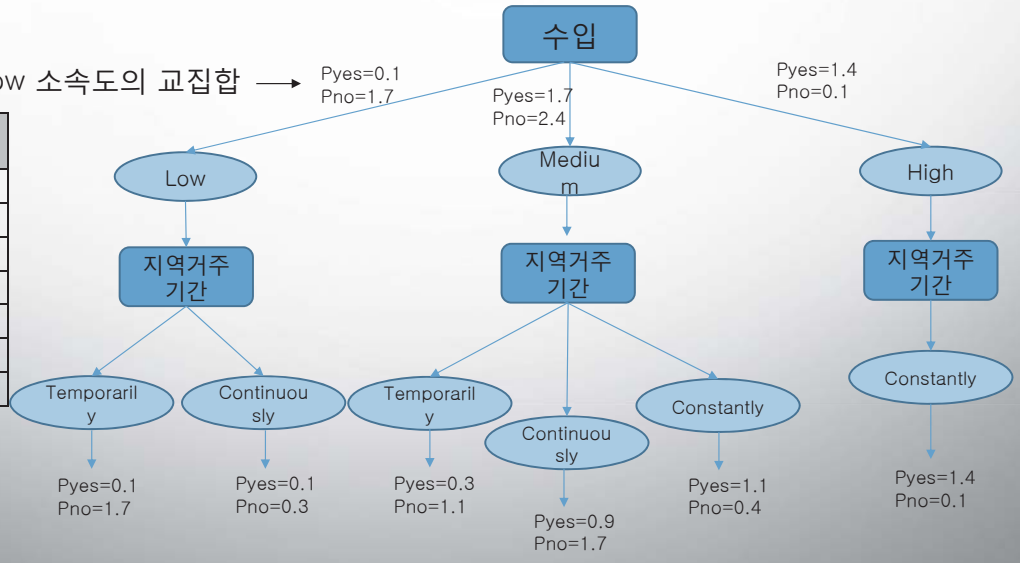
- 수입(Low,Medium,High)과 지역거주기간(Temp,Cont,Const)의 소속도의 교집합
- 소속도가 모두 0인 속성은 제거

수입	Low			Medium			High		
지역거주기간	Temporarily	Continuously	Constantly	Temporarily	Continuously	Constantly	Temporarily	Continuously	Constantly
D1	1	0	0	0	0	0	0	0	0
D2	0.6	0.2	0	0.4	0.2	0	0	0	0
D3	0.1	0.1	0	0.5	0.5	0	0	0	0
D4	0	0	0	0.2	0.8	0	0	0	0
D5	0	0	0	0	0.5	0.5	0	0	0
D6	0	0	0	0	0	0.6	0	0	0.4
D7	0	0	0	0	0	0	0	0	1

구현된 퍼지 의사결정트리의 모습

P_{yes}, P_{no}와 Low 소속도의 교집합 ⟶

P_{yes}	P_{no}	Low
0	1	1
0	1	0.6
0.1	0.9	0.1
0.3	0.7	0
0.7	0.3	0
0.9	0.1	0
1.0	0	0



- 노드 *[지역거주기간 = temporarily and 수입 = high]*
  [지역거주기간 = continuously and 수입 = high]
  [지역거주기간 = constantly and 수입 = low]
  들은 레코드의 소속도가 모두 0이므로 트리에서 제거

- 완성된 트리를 기반으로 지역거주기간이 25년이고 수입이 320만원인 신규 고객의 신용 등급을 정의

- 신규 고객은 두 영역에 속함: *[지역거주기간 = continuously and 수입 = medium]*
  [지역거주기간 = constantly and 수입 = high]
- 소속 함수에 따른 신규 고객의 소속도는 각각 [0.8, 0.8] / [0.2, 0.2] = $\mu_l(D_j)$
- 트리를 따라 얻은 P_{yes}와P_{no}의 값은 각각 [0.9, 1.7] 과 [1.4, 0.1] = P_k^l
- 신규 고객에 대한 신용도를 구하기 위해 아래와 같은 식을 사용

$$\sigma_j = \frac{\sum_l \sum_k P_k^l \cdot \mu_l(D_j) \cdot \chi_k}{\sum_l (\mu_l(D_j) \cdot \sum_k P_k^l)}, \quad (7)$$

$x_{yes} = 1,\ x_{no} = 0$

$$\therefore \frac{0.9 * 0.8 * \overset{x_{yes}}{1} + 1.7 * 0.8 * \overset{x_{no}}{0} + 1.4 * 0.2 * \overset{x_{yes}}{1} + 0.1 * 0.2 * \overset{x_{no}}{0}}{(0.9 + 1.7) * 0.8 + (1.4 + 0.1) * 0.2} = 0.420$$

- 신규 고객은 0.420 이라는 신용 등급을 받았으며, 대출을 받지 못함

결언

- 의사결정트리 모형은 데이터마이닝의 주요기법으로 자리잡고 있으며 SAS/EMINER, SPSS AnswerTree, R, Weka 등에서 이를 사용
- 의사결정트리는 예측하거나 분류하는 쉽고 투명한 방법으로, 데이터 구조에 대한 어떠한 가정을 갖지 않아, 다량의 샘플 필요
- 의사결정트리는 결과를 해석하고 이해하기에 쉽고, 자료를 가공할 필요가 거의 없는 방법
- 래덤 포레스트는 의사결정트리의 Overfitting 문제를 해결할 수 있으며, 산업체 등 현장에서의 활용도가 높음
- 최적의 결정 트리를 학습하는 문제는 NP-complete 문제로 알려져 있으며, 최적 결정 트리를 알아낸다고 보장 못함
- 퍼지의사결정트리는 각 특징값들이 경계면에 가까이 있는 경우 발생할 수 있는 분류 오류의 가능성을 줄일 수 있는 방법

감사합니다 ….

※ 저작권법에 의해 무단전재와 무단복재를 금합니다.
※ 본 권의 모든 저작권은 한국정보통신학회에 있습니다.

· 이 발표논문집은 2019년도 (재)부산인재평생교육진흥원의 지원을 받아 발간되었음.

한국정보통신학회 4차 산업 관련 지능형 기술 인력 양성 워크숍

서기 2019년 8월 23일 발행 발행처 : 사단법인 한국정보통신학회

부산광역시 부산진구 서면문화로 27, 1802호(부전동, 유원오피스텔)

전 화 : 051-463-3683

팩 스 : 051-464-3683 Email : kiice@kiice.org

홈페이지 : http://www.kiice.org

디지털 세상을 위한 아키텍쳐
휴인스가 함께합니다.

㈜휴인스는 ARM Core 기반의 솔루션, SoC 검증 플랫폼,
Drone 시스템,IoT 시스템, 안드로이드,임베디드 개발 솔루션,
코딩, 아두이노 시스템, 웨어러블 시스템, 로봇시스템분야의
전문 기술을 보유한 회사로서 지난 20여년간 중견기업으로 성장하였습니다.

AI Lab-OBJ

인공지능 딥러닝 사물인식 시스템

AI LaB-OBJ 자세히 보기 ▸

- 최적화된 Nvidia TX2 임베디드 시스템 사용
- GPU 프로그래밍 CUDA, 딥러닝을 위한 통계학 이해
- Ubuntu 16.04, CUDA, DNN, OpenCV 지원
- C, C++, Python 개발 언어 지원
- -Tensorflow, Caffe, Keras, Deep Learning Framework 지원
  MS-CNN, YOLO v3.0 Deep Learning Algorithm 지원
  YOLO를 이용한 실시간 객체감지 실습 지원

실습 과정

1. AI 개요
2. AI 관련 기술 정의
3,4. AI 최신 트렌드
5. 딥러닝이란?
6. 딥러닝 기술적 특징
7. 딥러닝 알고리즘의 종류
8. 딥러닝 사용 예
9. AI Lab-OBJ 소개
10. 숫자,알파벳 인식 알고리즘 (학습/실습)
11. 자동차 번호판 인식 알고리즘 (학습/실습)
12. 사물 인식 알고리즘 (학습/실습)
13. TX2 JetPack 설치
14. IoT 인터넷 제어 토이 실습

HUINS
Human Intelligent System

-본사
경기도 성남시 분당구 대왕판교로 670, 유스페이스2 B동 605호
대표전화 031-719-8200 **팩스** 031-719-8201 **제품구매 안내** sales@huins.com
-마곡연구소
서울특별시 강서구 마곡중앙8로3길 55
대표전화 02-3663-8201 **팩스** 02-3663-8207 **제품구매 안내** sales@huins.com

Communication & Network Service.
All that Network ! 포어링크 !

고객가치를 최우선으로 하는 ㈜포어링크는 2001년 NI 분야의 전문 기업으로 시작하여 네트워크 공사, 장비 설계/구축 및 유지보수 서비스를 기업체, 관공서, 교육기관, 병원 등의 고객사에게 제공하며 꾸준히 성장해 왔습니다.

네트워크 통합
Network Integration

네트워크 통합 서비스는
고객의 네트워크가 안정적으로
운영될 수 있도록 네트워크
컨설팅부터 네트워크 구축,
운영 서비스를 제공합니다.

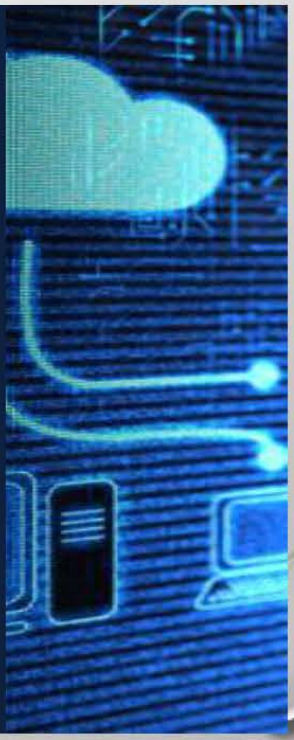
보안 및 IT기기
Security & IT Equipments

보안 서비스는
표준화된 보안 아키텍쳐를
수립하고 이에 기반한
고객의 정보시스템 환경에
부합하는 시스템을 권고,
통합 구축하는 서비스를
제공합니다.

Facility 구축 및 운영

전산실 구축, 이전 및 운영에
대한 풍부한 경험을 바탕으로
Facility 운영 전반에 걸친
Total 솔루션을 제공합니다.

유지보수 외..



CONTACT (HQ)

02) 2113- 9400

http://www.forelink.co.kr/

서울 금천구 가산디지털2로 184, 1308호

CONTACT (부산지사)

051) 504- 4343

부산 연제구 월드컵대로 83(KT&G빌딩 4F)

www.markany.com

CCTV관제센터의 **안전한 운영** 및 **체계적인 관리**를 위한

CCTV관제센터 최적화 통합 보안 솔루션

CCTV영상
반출보안솔루션

통합관제센터
자산관리솔루션

지능형선별관제솔루션

CCTV패스워드
보안관리솔루션

GIS솔루션

오남용감사
내부통제솔루션

Make IT Secure!

공공기관과 기업 모두에게 안전한 디지털 세상을 만들어 주는

Mark Any®

금융 IT Compliance 전문기업으로 Financial Information Technology의 새로운 눈이 되겠습니다.

Finance

금융과 기술의 융합!
완벽한 FINTech Technology
서비스의 중심! ITeyes

시스템을 바라보는
또 다른 새로운 눈
클라우드&빅데이터 서비스
Cloud&Bigdata

주식회사 아이티아이즈 www.iteyes.co.kr
정보시스템 개발 및 관리 서비스, 소프트웨어 솔루션 개발.도입 및 공급
서울특별시 영등포구 은행로 37 5층(여의도동, 기계진흥회관본관) **Tel.** 02-783-2970 / **Fax.** 02-783-5088

iT eyes Inc.

Since 1991 & future

www.hanscomic.com

hans com

SINCE 1991 주식회사한스콤정보통신

부산광역시 동래구 여고북로56 (사직3동 358-6 한빌딩 3층)

tel 051.507.0364 fax 051.507.0366

당신이 오늘 만날 수 있는 혁신

Innovation & Reality

4차산업 플랫폼 비즈니스 전문그룹

아이티센그룹 ITCENGROUP

아이티센 (KOSDAQ) | 소프트센 (KOSDAQ) | 굿센 | 큐시스템 (KONEX) | 콤텍시스템 (KOSPI) | 콤텍정보통신 (비상장) | 한국금거래소 (비상장)

Global IT Leader!

모든 비즈니스 영역을 통합하는 통찰력으로
고객의 니즈를 완벽히 분석한 최적의 서비스로
미래를 선도하는 최첨단 기술력으로

미래의 가치를 먼저 생각하는 기업

큰大믿을信
대신정보통신주식회사 Daishin Information & Communications Co., Ltd.
본사 : 광주광역시 서구 상무중앙로 110 Tel_062-225-7350 Fax_062-226-0716
서울 : 서울특별시 금천구 가산디지털1로, 205-28 대신정보통신빌딩 Tel_02-2017-5000 Fax_02-2107-5015
www.dsic.co.kr

KIICE 한국정보통신학회 사단법인 한국정보통신학회