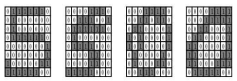
◆ 10-9평 36~39번

[36~39] 다음 글을 읽고 물음에 답하시오.

매일 쏟아지는 수많은 우편물들은 발송 지역별로 분류되어야 한다. 우편물 분류 작업은 우편번호 숫자를 인식함으로써 자동화될 수 있다. 이때 자동분류기는 환경과의 상호 작용에 기반한 경험적인 데이터로부터 스스로 성능을 향상시킬 수 있는 학습 능력을 갖춰야 한다. ㉠ 학습은 상호 작용의 정도에 따라 경험하는 데이터가 달라지고, 이러한 학습 데이터에 따라 자동분류기의 성능이 달라지게 된다. 즉, 자동분류기는 단순히 데이터를 기억하는 것이 아니라, 다양한 경험에서 새로운 정보를 추론하여 스스로 분류할 수 있는 능력을 갖춰야 한다.

	학습 데이터	실험 데이터
필기체 숫자		
입력 특징		
목표치	5 5 0 0	

우편번호 자동분류기가 학습하기 위해서는, 먼저 우편번호 숫자를 하나씩 분할하고, 0부터 9까지를 잘 구별할 수 있는 입력 특징을 찾아야 한다. 위 그림은 필기체 숫자를 가로, 세로 8등분하여 연필이 지나간 자리를 1, 그렇지 않으면 0의 값을 주어, 입력 특징을 추출한 것이다.

다음으로, 추출된 특징으로 학습할 때 분류기에 목표치를 제공함으로써 학습을 감독할 수 있다. 즉, 입력 특징에 대한 목표치가 제시되면 분류기는 데이터를 제시된 목표치로 분류하도록 학습한다. 이렇게 목표치를 이용하는 학습을 ㉡ 감독학습이라 한다. 숫자 분류기에 0부터 9까지 각각의 숫자에 대한 목표치가 제공되면, 분류기는 감독학습을 수행한다. 위의 그림에서 분류기는 네 개의 학습 데이터에 대한 입력 특징과 목표치를 통해 학습한다. 이 학습을 통해 두 개의 '5'와 두 개의 '0'을 각각 같은 숫자로 인식하면서, 동시에 '5'와 '0'을 서로 다른 숫자로 분류해 내는 함수를 만든다. 감독학습을 통해 올바르게 학습하였다면, 그림의 실험 데이터는 숫자 '5'로 인식된다.

그러면, 목표치를 주는 것이 어려운 경우에는 어떻게 학습할까? 목표치가 없을 때는 학습 데이터로 주어진 입력 특징들의 유사성을 찾아 군집화한다. 이와 같이 목표치가 제시되지 않는 학습을 무감독학습이라고 한다. 예컨대 위 그림에서 네 개의 필기체 숫자에 대한 입력 특징만 주어지면, 무감독학습은 비슷한 입력 특징을 가진 숫자들을 ⓐ 모아 '5' 또는 '0'에 대해 군집화하는 함수를 만든다. 무감독학습을 통해 올바르게 학습하였다면, 실험 데이터는 '5'의 군집과 유사한 것으로 인식된다.

이렇게 학습된 자동분류기는 실험 데이터를 정확하게 분류하였는지에 따라 그 성능이 평가된다. 이러한 과정을 통해 우편번호 자동분류기는 우편물을 지역별로 분류할 수 있게 된다.

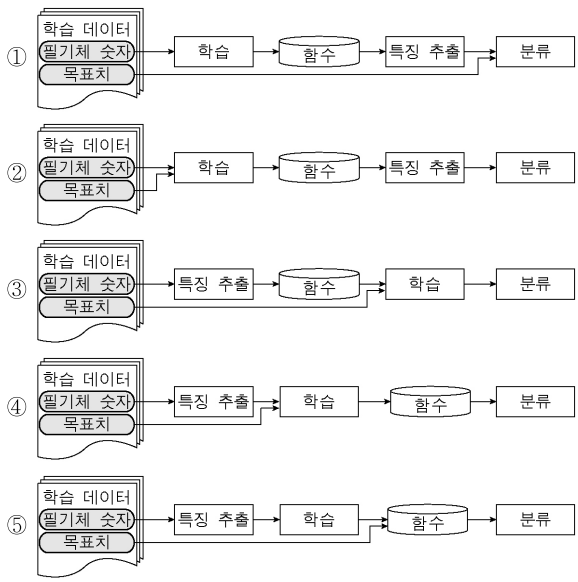
36. 위 글의 '우편번호 자동분류기'에 대한 설명으로 적절한 것은?

① 자동분류기의 성능은 학습 데이터의 양에 영향을 받지 않는다.
② 우리나라 우편번호 자동분류기는 총 6종류의 목표치를 이용한다.
③ 자동분류기의 학습은 일정한 종류의 필기체 숫자를 기억하는 것이다.
④ 자동분류기는 0부터 9까지의 차이를 최소화하는 입력 특징을 사용한다.
⑤ 자동분류기의 학습은 필기체 숫자의 목표치가 없으면, 유사한 입력 특징을 가진 것끼리 모은다.

37. 휴대 전화의 기능을 소개하는 문구 중, ㉠의 기능을 담은 예로 적절하지 않은 것은? [3점]

① 전화가 걸려 오면 등록된 수신 거부 목록과 일일이 대조하여, 목록에 있는 번호이면 수신을 거부한다.
② 휴대 전화를 든 손으로 등록된 단축 번호를 공중에 쓰면, 전화기가 숫자를 인식하여 자동으로 전화를 건다.
③ 사용자의 음성 특징을 추출하여 사용자와 타인의 음성을 분류하면, 사용자의 음성으로만 휴대 전화를 사용할 수 있다.
④ 휴대 전화에 닿는 형태를 유형화하여 접촉과 비접촉을 구별하면, 전화벨이 울리는 중에 휴대 전화에 손이 접촉할 경우 진동으로 전환된다.
⑤ 휴대 전화의 카메라로 촬영한 얼굴 영상들에서 색상값과 얼굴 형태 정보를 이용하여 얼굴과 얼굴이 아닌 것으로 분류하면, 사람이 움직여도 얼굴을 중심으로 촬영한다.

38. ㉡을 이용한 필기체 숫자 분류기의 구성도로 옳은 것은?



39. 문맥상 ⓐ와 바꾸어 쓸 수 있는 한자어로 가장 적절한 것은?

① 취합(聚合)하여 ② 융합(融合)하여
③ 조합(組合)하여 ④ 규합(糾合)하여
⑤ 결합(結合)하여

- 정답: 36.⑤ 37.① 38.④ 39.①

◆ 17-6평 16~19번

[16~19] 다음 글을 읽고 물음에 답하시오.

인간의 신경 조직을 수학적으로 모델링하여 컴퓨터가 인간처럼 기억·학습·판단할 수 있도록 구현한 것이 인공 신경망 기술이다. 신경 조직의 기본 단위는 뉴런인데, ⓐ 인공 신경망에서는 뉴런의 기능을 수학적으로 모델링한 퍼셉트론을 기본 단위로 사용한다.

ⓑ 퍼셉트론은 입력값들을 받아들이는 여러 개의 ⓒ 입력 단자와 이 값을 처리하는 부분, 처리된 값을 내보내는 한 개의 출력 단자로 구성되어 있다. 퍼셉트론은 각각의 입력 단자에 할당된 ⓓ 가중치를 입력값에 곱한 값들을 모두 합하여 가중합을 구한 후, 고정된 ⓔ 임계치보다 가중합이 작으면 0, 그렇지 않으면 1과 같은 방식으로 ⓕ 출력값을 내보낸다.

이러한 퍼셉트론은 출력값에 따라 두 가지로만 구분하여 입력값들을 판정할 수 있을 뿐이다. 이에 비해 복잡한 판정을 할 수 있는 인공 신경망은 다수의 퍼셉트론을 여러 계층으로 배열하여 한 계층에서 출력된 신호가 다음 계층에 있는 모든 퍼셉트론의 입력 단자에 입력값으로 입력되는 구조로 이루어진다. 이러한 인공 신경망에서 가장 처음에 입력값을 받아들이는 퍼셉트론들을 입력층, 가장 마지막에 있는 퍼셉트론들을 출력층이라고 한다.

㉠ 어떤 사진 속 물체의 색깔과 형태로부터 그 물체가 사과인지 아닌지를 구별할 수 있도록 인공 신경망을 학습시키는 경우를 생각해 보자. 먼저 학습을 위한 입력값들 즉 학습 데이터를 만들어야 한다. 학습 데이터를 만들기 위해서는 사과 사진을 준비하고 사진에 나타난 특징인 색깔과 형태를 수치화해야 한다. 이 경우 색깔과 형태라는 두 범주를 수치화하여 하나의 학습 데이터로 묶은 다음, '정답'에 해당하는 값과 함께 학습 데이터를 인공 신경망에 제공한다. 이때 같은 범주에 속하는 입력값은 동일한 입력 단자를 통해 들어가도록 해야 한다. 그리고 사과 사진에 대한 학습 데이터를 만들 때에 정답인 '사과이다'에 해당하는 값을 '1'로 설정하였다면 출력값 '0'은 '사과가 아니다'를 의미하게 된다.

인공 신경망의 작동은 크게 학습 단계와 판정 단계로 나뉜다. 학습 단계는 학습 데이터를 입력층의 입력 단자에 넣어 주고 출력층의 출력값을 구한 후, 이 출력값과 정답에 해당하는 값의 차이가 줄어들도록 가중치를 갱신하는 과정이다. 어떤 학습 데이터가 주어지면 이때의 출력값을 구하고 학습 데이터와 함께 제공된 정답에 해당하는 값에서 출력값을 뺀 값 즉 오차 값을 구한다. 이 오차 값의 일부가 출력층의 출력 단자에서 입력층의 입력 단자 방향으로 되돌아가면서 각 계층의 퍼셉트론별로 출력 신호를 만드는 데 관여한 모든 가중치들에 더해지는 방식으로 가중치들이 갱신된다. 이러한 과정을 다양한 학습 데이터에 대하여 반복하면 출력값들이 각각의 정답 값에 수렴하게 되고 판정 성능이 좋아진다. 오차 값이 0에 근접하게 되거나 가중치의 갱신이 더 이상 이루어지지 않게 되면 학습 단계를 마치고 판정 단계로 전환한다. 이때 판정의 오류를 줄이기 위해서는 학습 단계에서 대상들의 변별적 특징이 잘 반영되어 있는 서로 다른 학습 데이터를 사용하는 것이 좋다.

16. 윗글에 따를 때, ⓐ~ⓕ에 대한 설명으로 적절하지 않은 것은?

① ⓑ는 ⓐ의 기본 단위이다.
② ⓒ는 ⓑ를 구성하는 요소 중 하나이다.
③ ⓓ가 변하면 ⓔ도 따라서 변한다.
④ ⓔ는 ⓕ를 결정하는 기준이 된다.
⑤ ⓐ가 학습하는 과정에서 ⓕ는 ⓓ의 변화에 영향을 미친다.

17. 윗글에 대한 이해로 적절하지 않은 것은?

① 퍼셉트론의 출력 단자는 하나이다.
② 출력층의 출력값이 정답에 해당하는 값과 같으면 오차 값은 0이다.
③ 입력층 퍼셉트론에서 출력된 신호는 다음 계층 퍼셉트론의 입력값이 된다.
④ 퍼셉트론은 인간의 신경 조직의 기본 단위의 기능을 수학적으로 모델링한 것이다.
⑤ 가중치의 갱신은 입력층의 입력 단자에서 출력층의 출력 단자 방향으로 진행된다.

18. 윗글을 바탕으로 ㉠에 대해 추론한 것으로 적절하지 않은 것은?

① 학습 데이터를 만들 때는 색깔이나 형태가 다른 사과의 사진을 선택하는 것이 좋겠군.
② 학습 데이터에 두 가지 범주가 제시되었으므로 입력층의 퍼셉트론은 두 개의 입력 단자를 사용하겠군.
③ 색깔에 해당하는 범주와 형태에 해당하는 범주를 분리하여 각각 서로 다른 학습 데이터로 만들어야 하겠군.
④ 가중치가 더 이상 변하지 않는 단계에 이르면 '사과'인지 아닌지를 구별하는 학습 단계가 끝났다고 볼 수 있겠군.
⑤ 학습 데이터를 만들 때 사과 사진의 정답에 해당하는 값을 0으로 설정하였다면, 출력층의 출력 단자에서 0 신호가 출력되면 '사과이다'로, 1 신호가 출력되면 '사과가 아니다'로 해석해야 되겠군.

19. 윗글을 바탕으로 <보기>를 이해한 내용으로 가장 적절한 것은? [3점]

<보 기>

아래의 [A]와 같은 하나의 퍼셉트론을 [B]를 이용해 학습시키고자 한다.

[A]
○ 입력 단자는 세 개(a, b, c)
○ a, b, c의 현재의 가중치는 각각 W_a=0.5, W_b=0.5, W_c=0.1
○ 가중합이 임계치 1보다 작으면 0을, 그렇지 않으면 1을 출력

[B]
○ a, b, c로 입력되는 학습 데이터는 각각 I_a=1, I_b=0, I_c=1
○ 학습 데이터와 함께 제공되는 정답=1

① [B]로 학습시키기 위해서는 판정 단계를 먼저 거쳐야 하겠군.
② 이 퍼셉트론이 1을 출력한다면, 가중합이 1보다 작았기 때문이겠군.
③ [B]로 한 번 학습시키고 나면 가중치 W_a, W_b, W_c가 모두 늘어나 있겠군.
④ [B]로 여러 차례 반복해서 학습시키면 퍼셉트론의 출력값은 0에 수렴하겠군.
⑤ [B]의 학습 데이터를 한 번 입력했을 때 그에 대한 퍼셉트론의 출력값은 1이겠군.

- 정답: 16.③ 17.⑤ 18.③ 19.③

◆ 25 수능 10~13번

[10~13] 다음 글을 읽고 물음에 답하시오.

문장이나 영상, 음성을 만들어 내는 인공 지능 생성 모델 중 확산 모델은 영상의 복원, 생성 및 변환에 뛰어난 성능을 보인다. 확산 모델의 기본 발상은, 원본 이미지에 노이즈를 점진적으로 추가하였다가 그 노이즈를 다시 제거해 나가면 원본 이미지를 복원할 수 있다는 것이다. 노이즈는 불필요하거나 원하지 않는 값을 의미한다. 원하는 값만 들어 있는 원본 이미지에 노이즈를 단계별로 더하면 노이즈가 포함된 확산 이미지가 되고, 여러 단계를 거치면 결국 원본 이미지가 어떤 이미지였는지 전혀 알아볼 수 없는 노이즈 이미지가 된다. 역으로, 단계별로 더해진 노이즈를 알 수 있다면 노이즈 이미지에서 원본 이미지를 복원할 수 있다. 확산 모델은 노이즈 생성기, 이미지 연산기, 노이즈 예측기로 구성되며, 순확산 과정과 역확산 과정 순으로 작동한다.

순확산 과정은 이미지에 노이즈를 추가하면서 노이즈 예측기를 학습시키는 과정이다. 첫 단계에서는, 노이즈 생성기에서 노이즈를 만든 후 이미지 연산기가 이 노이즈를 원본 이미지에 더해서 노이즈가 포함된 확산 이미지를 출력한다. 다음 단계부터는 노이즈 생성기에서 만든 노이즈를 이전 단계에서 출력된 확산 이미지에 더한다. 이러한 단계를 충분히 반복하면 최종적으로 노이즈 이미지가 출력된다. 이때 더해지는 노이즈는 크기나 분포 양상 등 그 특성이 단계별로 다르다. 따라서 노이즈 예측기는 단계별로 확산 이미지를 입력받아 이미지에 포함된 노이즈의 특성을 추출하여 수치들로 표현하고, 이 수치들을 바탕으로 노이즈를 예측한다. 노이즈 예측기 내부의 이러한 수치들을 잠재 표현이라고 한다. 노이즈 예측기는 잠재 표현을 구하고 노이즈를 예측하는 방식을 학습한다.

노이즈 예측기의 학습 방법은 기계 학습 중에서 지도 학습에 해당한다. 지도 학습은 학습 데이터에 정답이 주어져 출력과 정답의 차이가 작아지도록 모델을 학습시키는 방법이다. 노이즈 예측기를 학습시킬 때는 노이즈 생성기에서 만들어 넣어 준 노이즈가 정답에 해당하며 이 노이즈와 예측된 노이즈 사이의 차이가 작아지도록 학습시킨다.

역확산 과정은 노이즈 이미지에서 노이즈를 제거하여 원본 이미지를 복원하는 과정이다. 노이즈를 제거하려면 이미지에 단계별로 어떤 특성의 노이즈가 더해졌는지 알아야 하는데 노이즈 예측기가 이 역할을 한다. 노이즈 이미지 또는 중간 단계에서의 확산 이미지를 노이즈 예측기에 입력하면 이미지에 포함된 노이즈의 특성을 추출하여 잠재 표현을 구하고 이를 바탕으로 노이즈를 예측한다. 이미지 연산기는 입력된 확산 이미지로부터 이 노이즈를 빼서 현 단계의 노이즈를 제거한 확산 이미지를 출력한다. 확산 이미지에 이런 단계를 반복하면 결국 노이즈가 대부분 제거되어 원본 이미지에 가까운 이미지만 남게 된다.

한편, 많은 종류의 이미지를 학습시킨 후 학습된 이미지의 잠재 표현에 고유 번호를 붙이면 역확산 과정에서 이미지를 선택하여 생성할 수 있다. 또한 잠재 표현의 수치들을 조정하면 다른 특성의 노이즈가 생성되어 여러 이미지를 혼합하거나 실재하지 않는 이미지를 만들어 낼 수도 있다.

10. 학생이 윗글을 읽은 방법으로 적절하지 <u>않은</u> 것은?

① 확산 모델이 지도 학습을 사용한다는 점에 주목하고, 지도 학습 방법이 확산 모델에 어떻게 적용되는지 확인하며 읽었다.
② 확산 모델이 두 가지 과정으로 이루어진다는 점에 주목하고, 두 과정 중 어느 과정이 선행되어야 하는지 살피며 읽었다.
③ 확산 모델에서 노이즈의 중요성을 파악하고, 사용되는 노이즈의 종류가 모델의 성능에 미치는 영향을 이해하며 읽었다.
④ 잠재 표현의 개념을 파악하고, 그 개념을 바탕으로 확산 모델이 노이즈를 예측하고 제거하는 원리를 이해하며 읽었다.
⑤ 확산 모델의 구성 요소를 파악하고, 그 구성 요소가 노이즈 처리 과정에서 어떤 기능을 하는지 확인하며 읽었다.

11. 윗글을 이해한 내용으로 가장 적절한 것은?

① 노이즈 생성기는 순확산 과정에서만 작동한다.
② 확산 모델에서의 학습은 역확산 과정에서 이루어진다.
③ 이미지 연산기와 노이즈 예측기는 모두 확산 이미지를 출력한다.
④ 노이즈 예측기를 학습시킬 때는 예측된 노이즈가 정답으로 사용된다.
⑤ 역확산 과정에서 단계가 반복될수록 출력되는 확산 이미지는 원본 이미지와의 유사성이 줄어든다.

12. 잠재 표현에 대한 설명으로 적절하지 <u>않은</u> 것은?

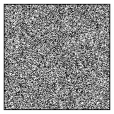
① 잠재 표현의 수치들을 조정하면 여러 이미지를 혼합할 수 있다.
② 역확산 과정에서 잠재 표현이 다르면 예측되는 노이즈가 다르다.
③ 확산 모델의 학습에는 잠재 표현을 구하는 방식이 포함되어 있다.
④ 잠재 표현은 이미지에 더해진 노이즈의 크기나 분포 양상에 따라 다른 값들이 얻어진다.
⑤ 잠재 표현은 노이즈 예측기가 원본 이미지를 입력받아 노이즈의 특성을 추출한 결과이다.

13. 윗글을 바탕으로 <보기>를 이해한 내용으로 적절하지 않은 것은? [3점]

<보 기>

A 단계는 확산 모델 과정 중 한 단계이다. ㉠은 원본 이미지이고, ㉡은 확산 이미지 중의 하나이며, ㉢은 노이즈 이미지이다. (가)는 이미지가 A 단계로 입력되는 부분이고, (나)는 이미지가 A 단계에서 출력되는 부분이다.

(가) ⇨ [A 단계] ⇨ (나)

㉠  ㉡  ㉢ 

① (가)에 ㉠이 입력된다면, A 단계의 이미지 연산기에서는 ㉠에 노이즈를 더하겠군.

② (나)에 ㉢이 출력된다면, A 단계의 노이즈 생성기에서 생성된 노이즈가 이미지 연산기에서 확산 이미지에 더해졌겠군.

③ 순확산 과정에서 (가)에 ㉡이 입력된다면, A 단계의 노이즈 예측기에서 예측한 노이즈가 이미지 연산기에 입력되겠군.

④ 역확산 과정에서 (가)에 ㉢이 입력된다면, A 단계의 이미지 연산기에서는 ㉢에서 노이즈를 빼겠군.

⑤ 역확산 과정에서 (나)에 ㉡이 출력된다면, A 단계의 노이즈 예측기에서 예측한 노이즈가 이미지 연산기에 입력되었겠군.

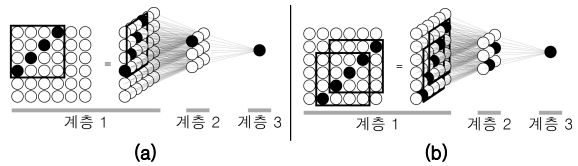
- 정답: 10.③ 11.① 12.⑤ 13.③

◆ 18년 3월 고2 24~28번

[24～28] 다음 글을 읽고 물음에 답하시오.

빛은 망막의 광수용기 세포에서 수용되어 전기 신호로 변환된 뒤, 뇌의 시각 피질로 전달된다. ㉠ 후벨과 위젤은 망막에 비춰진 빛에 대해 고양이의 시각 피질 세포가 어떻게 반응하는지 실험하였다. 그들은 이를 통해 시각 피질 세포가 망막의 일정 영역 내 광수용기 세포들과 연결되어 있다는 사실을 알아냈다. 하나의 시각 피질 세포와 연결된 망막상의 일정 영역을 해당 시각 피질 세포의 '수용장'이라고 한다.

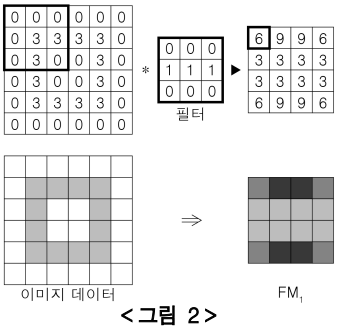
또한 이 실험을 통해 시각 피질이 하위의 '단순 세포'와 상위의 '복잡 세포'의 다층 구조로 구성되어 있다는 점이 밝혀졌다. 단순 세포와 복잡 세포 모두 각각의 수용장에 비친 특정한 각도를 가진 선분 모양의 빛에 활성화된다. 하지만 단순 세포가 수용장 내 특정 위치의 빛에만 활성화되는데 반해, 복잡 세포는 수용장이 단순 세포보다 넓고, 수용장에 비춰진 빛의 위치 변화에 관계없이 활성화된다. 이는 복잡 세포가 다수의 단순 세포들로부터 전기 신호를 전달받아 활성화되기 때문이다.



<그림 1>

<그림 1>은 이러한 시각 피질 세포들의 전기 신호 전달 과정을 다층 모형으로 나타낸 것이다. 모형의 각 층은 유닛들로 구성되는데, 계층 1의 각 유닛은 망막의 광수용기 세포에, 계층 2의 각 유닛은 단순 세포에, 계층 3의 유닛은 복잡 세포에 대응된다. 이때, 검은색 유닛은 해당 유닛이 활성화되었음을 의미하며, 계층 1의 사각형 영역은 계층 2의 활성화된 유닛의 수용장을 표시한 것이다. (a)와 (b)는 각각의 사선 패턴의 위치에 따른 각 유닛들의 활성화 상태를 나타낸 것이다. 계층 2의 각 유닛은 자신의 수용장 안의 특정한 위치에 특정한 각도의 사선 패턴이 입력되면 활성화된다. 계층 3의 유닛은 계층 2의 유닛 중에 하나라도 활성화되면 활성화된다.

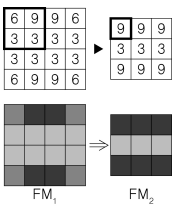
'합성곱 신경망'은 이미지 인식(image recognition)*을 위해 만들어진 인공 신경망으로서, <그림 1>과 같은 다층 구조의 신경망 모형을 수학적으로 구조화한 것이다. 합성곱 신경망은 '합성곱층'과 '통합층'으로 구성되며, 이들은 각각 합성곱 연산과 통합 연산에 의해 출력된다. 먼저, 합성곱 연산은 특정한 크기의 필터가 이미지 데이터의 왼쪽 상단에서 오른쪽 하단까지 일정 간격으로 이동해 가며 이미지 데이터와 필터의 곱을 합산하는 과정이다. 이때 필터는 이미지 데이터의 국부 영역에 존재하는 특정한 기하학적 패턴을 검출하는 역할을 한다.



<그림 2>

예를 들어, <그림 2>와 같이 '□'의 형태를 가진 6×6 크기의 이미지 데이터로부터 수평 방향의 패턴을 추출한다고 하자. 이때, 각 유닛의 숫자는 명암을 0부터 10까지의 수치로 나타낸 것이다. 필터의 크기가 3×3이고 이동 간격을 1 유닛 단위로 설정했다면, 필터가 왼쪽 상단에서 오른쪽 하단으로 한 칸씩 이동해 가면서 합성곱을 16번 연산하고 4×4 크기의 '특징 지도'(feature map, FM)가 출력된다. <그림 2>에서 특징 지도 FM_1의 가장 왼쪽 위 유닛 값 '6'은 이미지 데이터의 왼쪽 위 3×3의 영역과 필터와의 곱의 총합인 '0×0 + 0×0 + 0×0 + 0×1 + 3×1 + 3×1 + 0×0 + 3×0 + 0×0'의 연산을 통해 구해진 것이다.

이렇게 필터를 이용해 이미지 데이터에 합성곱 연산을 수행하면 필터의 특성에 맞게 강조된 특징 지도를 얻을 수 있다. <그림 2>는 합성곱 연산 결과 수평 방향의 패턴이 강조되고 데이터 크기는 6×6에서 4×4로 줄어 출력된 특징 지도를 보여 준다. 이때, 필터의 이동 간격이 크게 설정된다면 출력되는 특징 지도의 크기를 줄여 데이터 처리를 빠르게 할 수 있는 장점이 있지만, 이미지의 특징을 놓칠 가능성이 증가하게 되는 단점이 있다.



<그림 3>

다음으로, 통합 연산은 합성곱층의 일정 범위 안에 있는 유닛 값들을 정해진 규칙에 따라 하나의 값으로 통합하는 연산이다. 통합 연산 규칙에는 최댓값 통합 규칙, 평균값 통합 규칙 등 여러 종류가 있는데, 이를 통해 새롭게 출력된 특징 지도로 통합층이 구성된다. <그림 3>은 <그림 2>의 FM_1을 2×2 범위로 최댓값 통합 규칙에 따라 통합 연산한 것이다. 이때, 통합 연산의 범위를 왼쪽 상단에서 오른쪽 하단까지 1 유닛 단위로 이동하도록 설정하면 3×3 크기의 새로운 특징 지도 FM_2가 출력된다.

[가] 합성곱 연산을 통해 이미지의 어떤 영역에 어떤 패턴이 있는지를 추출할 수 있으며, 다양한 필터를 통해 이를 반복하면 이미지 속 사물을 인식할 수 있다. 하지만 연산을 반복하는 과정에서 패턴의 위치 정보를 계속 유지하게 되는데, 이는 일반적으로 불필요한 정보이다. 왜냐하면, 합성곱 연산을 통해 출력된 특징 지도 내에서 서로 인접한 유닛들은 미세한 위치 정보만 다를 뿐, 거의 비슷한 패턴 정보를 담고 있기 때문이다. 이때, 통합 연산 수행은 합성곱 연산의 결과에서 위치 정보를 줄여 주는 역할을 한다.

합성곱 연산과 통합 연산을 통해 위치 정보는 축약되고 패턴 정보는 강조된 특징 지도가 출력된다. 그리고 이 특징 지도를 인공 지능 네트워크인 '전체 연결층'에 입력하여 이미지 인

식 결과를 출력할 수 있다. 또한 입력된 이미지가 많아질수록 인공 신경망의 기계 학습을 통해 합성곱 신경망이 스스로 필터의 수치를 갱신함으로써 이미지 인식의 정확성이 높아지게 된다. 하지만 합성곱 연산 및 통합 연산의 횟수, 필터의 크기 및 이동 간격, 통합 연산 규칙 등은 초기 설정 값이 계속 유지되므로 이를 고려하여 합성곱 신경망을 설계해야 한다. 최근 인공 지능 기술이 발전함에 따라 합성곱 신경망은 사진 자동 분류, 필기 인식 등 다양한 영역으로 확장되고 있다.

* 이미지 인식 : 이미지 속 사물이 무엇인지를 알아내는 것.

24. 윗글에 대한 이해로 적절하지 <u>않은</u> 것은?

① 통합 연산은 합성곱층의 일정 범위 내의 값들을 하나의 값으로 통합하는 기능을 한다.
② 시각 피질의 복잡 세포는 단순 세포로부터 전달받은 전기 신호를 전체 연결층에 전달한다.
③ 시각 피질의 단순 세포는 수용장 내에 비춰진 특정 각도의 선분 모양의 빛에 활성화된다.
④ 합성곱 신경망으로 이미지를 인식하려면 특징 지도에 특정 패턴에 대한 정보가 담겨 있어야 한다.
⑤ 합성곱 신경망은 합성곱 연산과 통합 연산을 통해 이미지의 패턴 정보가 강조된 특징 지도를 추출한다.

25. <보기>는 ㉠을 재구성한 실험에 대한 설명이다. <보기>와 윗글의 <그림 1>을 이해한 것으로 적절하지 <u>않은</u> 것은? [3점]

< 보 기 >

다양한 빛 자극에 대해 시각 피질 세포가 어떻게 반응하는지 알기 위해, 선분 모양의 빛을 고양이의 망막에 비춘다.

이때, 빛의 각도는 각도 Ⓐ와 Ⓑ로, 빛이 비추어지는 수용장 내 위치는 위치 ㉮와 ㉯로 각각 다르게 한다. 그 결과 세포 A와 B는 서로 다른 반응을 보였다.

(단, 세포 A와 B는 서로 다른 시각 피질 세포이며, 망막의 특정 영역을 수용장으로 공유한다.)

실험				실험 결과	
	빛의 각도	빛의 위치		세포 A	세포 B
자극 1	Ⓐ	㉮	⇨	○	○
자극 2	Ⓐ	㉯	⇨	○	×
자극 3	Ⓑ	㉮	⇨	×	×
자극 4	Ⓑ	㉯	⇨	×	×

(○ : 활성화, × : 비활성화)

① '자극 1'의 실험 결과를 고려하면, '세포 A'와 '세포 B'가 반응하는 빛의 각도는 같겠군.
② '자극 1'과 '자극 2'의 실험 결과를 고려하면, '세포 A'의 수용장이 '세포 B'의 수용장보다 더 넓겠군.
③ '자극 1'과 '자극 3'의 실험 결과를 비교하면, '세포 A'는 각도 Ⓑ의 빛에는 반응하지 않겠군.
④ '세포 A'는 <그림 1>의 '계층 3'의 유닛에, '세포 B'는 '계층 2'의 유닛에 해당되겠군.
⑤ '자극 1'과 '자극 2'의 실험 결과는 <그림 1>의 (a)에, '자극 3'과 '자극 4'의 실험 결과는 (b)에 해당되겠군.

26. 필터에 대한 이해로 적절하지 <u>않은</u> 것은?

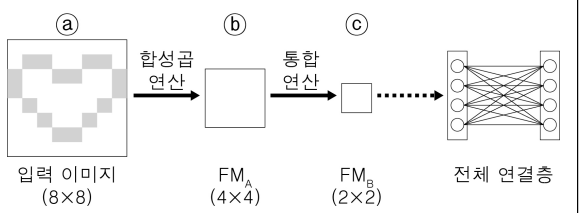
① 합성곱 연산을 수행하면 필터의 특성이 반영된 특징 지도가 출력된다.
② 필터의 기능은 이미지 데이터에서 특정한 기하학적 패턴을 검출하는 것이다.
③ 적절한 필터를 통해 합성곱 연산을 반복하여 이미지 속 사물을 인식할 수 있다.
④ 필터의 크기와 이동 간격의 비율은 합성곱 신경망에 의해 자동적으로 변화된다.
⑤ 필터의 매개를 통해 이미지 속 사물의 패턴에 대한 정보가 합성곱층에 반영된다.

27. [가]를 고려할 때, '통합 연산'을 수행하는 이유로 적절한 것은?

① 통합 연산 수행 이전과 이후, 이미지 속 사물에 대한 인식 결과가 달라지기 때문이다.
② 통합층의 각 유닛에 담긴 정보는 합성곱층의 각 유닛에 담긴 정보와 관련이 없기 때문이다.
③ 이미지 속 사물의 위치 정보를 표시하기 위한 추가적인 합성곱 연산이 필요하기 때문이다.
④ 합성곱 연산을 수행한 결과에 이미지 인식에는 불필요한 위치 정보가 포함되어 있기 때문이다.
⑤ 통합 연산은 합성곱층에 포함된 이미지 속 사물의 패턴 정보를 추출하는 역할을 하기 때문이다.

28. <보기>는 '♡' 모양의 디지털 이미지를 인식하는 과정의 일부를 도식화한 것이다. 이에 대해 이해한 것으로 적절하지 <u>않은</u> 것은?

< 보 기 >



(단, 유닛 한 단위의 데이터 크기는 동일하다.)

① ⓑ의 데이터 크기는 ⓐ에 비해 작겠군.
② 필터의 이동 간격을 1 유닛 단위로 설정했다면 ⓑ를 출력하기 위해 5×5 필터가 사용되었겠군.
③ 2×2 범위로 평균값 통합을 통해 ⓒ를 출력했다면, ⓒ의 데이터 크기는 ⓑ의 25%로 감소하였겠군.
④ 2×2 범위로 최댓값 통합 규칙을 사용하여 ⓑ를 통합 연산한 경우, 해당 범위의 유닛 값들 중 최댓값이 ⓒ의 하나의 유닛 값으로 도출되겠군.
⑤ ⓑ에서 ⓒ를 출력하기 위한 통합 연산에는 '♡' 모양의 특징을 검출할 수 있는 필터가 적용되었겠군.

- 출전: 오카타니 타케유키 저· 심효섭 역. 《딥 러닝 제대로 시작하기》. (제이펍, 2016)
- 정답: 24.② 25.⑤ 26.④ 27.④ 28.⑤

[22~24] 다음 글을 읽고 물음에 답하시오.

1990년대 이후 온톨로지(ontology)는 인공지능 연구에서 각광을 받고 있다. 연구자들마다 '온톨로지'란 용어를 조금씩 다른 의미로 사용하고 있지만, 널리 받아들여지는 정의는 "관심 영역 내 공유된 개념화에 대한 형식적이고 명시적인 명세"다. 여기서 '관심 영역'은 특정 영역 중심적이라는 것을, '공유된'은 관련된 사람들의 합의에 의한 것이라는 것을, '개념화'는 현실 세계에 대한 모형이라는 것을 뜻한다. 즉 특정 영역의 지식을 모델링하여 구성원들의 지식 공유 및 재사용을 가능하게 하는 것이 바로 온톨로지인 것이다. 또 '형식적'은 기계가 읽고 처리할 수 있는 형태로 온톨로지를 표현해야 한다는 것을 뜻한다. 그 결과로서 얻어지는 '명시적인 명세'는 일종의 공학적 구조물로서 다양한 용도로 사용된다.

온톨로지를 사전과 비교하면 '개념화'를 쉽게 이해할 수 있다. 사전에는 각각의 표제어에 대해 뜻풀이, 동의어, 반대어 등 언어적 특성들이 정리되어 있다. 온톨로지에는 표제어 대신 개념이, 그리고 언어적 특성들 대신 개념들 간 논리적 특성들이 기록된다. '개념(class)'은 어떤 공통된 속성들을 공유하는 '개체들(instances)'의 집합이고, 개체는 세상에 존재하는 구체적인 개별자이다. 온톨로지에서 개념은 관계를 통해 다른 개념들과 연결된다. 필수적인 관계는 개념 간의 계층 구조를 형성하는 상속 관계이다. 상속 관계에서 하위 개념은 상위 개념의 모든 속성을 물려받는다. 예컨대 '스누피'라는 특정 개체가 속한 견종 '몰티즈'라는 개념은 '개'의 하위 개념이므로, '몰티즈'는 상위 개념인 '개'가 가진 모든 속성을 물려받는다. 널리 사용되는 또 다른 관계로 부분-전체 관계가 있다. 이외에도 온톨로지에는 관계를 포함한 다양한 논리적 특성들을 기록할 수 있다.

온톨로지 표현 언어는 대부분 일차 술어 논리에 기초를 두고 있다. 일차 술어 논리는 '모든'과 '어떤'을 변수와 함께 사용하는 언어로 표현력이 매우 뛰어나다. 예컨대 "진짜 이탈리아 피자는 오직 얇고 바삭한 베이스만을 갖는다."를 일차 술어 논리로 옮기면 "모든 x에 대해, 만약 x가 진짜 이탈리아 피자라면, 얇고 바삭한 베이스인 어떤 y가 존재하고 x는 y를 베이스로 갖는다."가 된다. 그런데 이것이 반드시 장점인 것은 아니다. 일차 술어 논리로 정교하고 복잡하게 표현된 온톨로지를 막상 기계는 효율적으로 다룰 수 없는 경우가 발생하기 때문이다. 따라서 온톨로지 표현 언어는 일차 술어 논리에 각종 제약을 두어 표현력을 줄이는 대신 취급을 용이하도록 한 것이 대부분이다. 예컨대 월드 와이드 웹 컨소시움의 권고안인 '웹 온톨로지 언어' OWL에는 Lite, DL, Full의 세 가지 버전이 있는데, 후자로 갈수록 표현력이 커진다. 즉 OWL DL은 OWL Lite의 확장이고 OWL Full은 OWL DL의 확장이다. OWL DL까지는 계산학적 완전성과 결정 가능성이 보장된다. 이는 OWL DL로 표현된 온톨로지에서는 추론 엔진이 유한한 시간 내에 항상 해를 찾을 수 있음을 뜻한다.

OWL을 쓰면 복잡하고 다양한 논리적 특성들을 표현할 수 있지만 논리학에 익숙하지 않은 사용자에게 OWL은 너무 어렵다. 이로 인해 그 이름과는 달리, 웹에서 OWL이 널리 쓰이는 것은 아직까지 요원해 보인다. 오히려 전문 지식에 대한 정교한 논리적

표현이 요구되는 영역에서는 OWL이 이용되는 경우가 있다. 예컨대 미국 국립암센터에서 개발한 의료 영역 온톨로지인 NCI 시소러스는 OWL 포맷으로도 제공되는데, 이것은 약 4만 개의 개념과 백 개 이상의 관계로 이루어져 있다. 이외에도 의료 영역은 일찍부터 여러 그룹에서 각기 목적에 맞는 온톨로지를 발전시켜 왔다. 대표적인 것으로는 UMLS, SNOMED-CT 등이 있다.

온톨로지는 일반적으로 특정 영역 종사자들의 관심과 필요에 의해 구축되나 반드시 그런 것은 아니다. 1984년 개발이 시작된 Cyc는 인간의 모든 지식을 담고자 하는 대규모 온톨로지다. 지식공학자 소와(Sowa)는 철학의 연구 성과를 적극적으로 수용한 상위 수준 온톨로지를 제시한 바 있다. 세상에 존재하는 모든 것을 분류하려면 시간, 공간과 같은 일반적인 개념들을 다루어야만 하는데, 이는 철학자들이 이런 개념들에 대해 가장 오랫동안 깊이 사유했기 때문이다.

22. 온톨로지에 대한 설명으로 적절하지 않은 것은?

① 지식의 공유와 재사용을 위해 설계된 인공물이다.
② 대상 체계의 개념 구조를 명시적으로 드러내고자 한다.
③ 실제 사용되려면 기계가 처리할 수 있는 형태로 표현되어야 한다.
④ 개념과 그 개념에 속한 개체들은 상속 관계에 의해 서로 연결된다.
⑤ 동일한 영역에서도 종사자들의 관심과 필요에 따라 서로 다른 온톨로지가 구축될 수 있다.

23. 온톨로지 표현 언어에 대해 추론한 내용으로 적절한 것만을 <보기>에서 있는 대로 고른 것은?

<보 기>

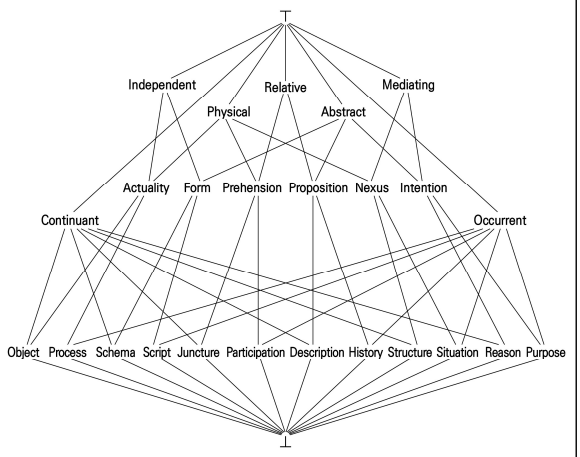
ㄱ. 동일한 온톨로지를 서로 다른 두 개의 언어로 각각 표현하기 위해서는 이들 언어의 표현력이 동등해야 한다.
ㄴ. 일차 술어 논리 표현 "모든 x에 대해, x가 빵이면 x는 장미이다."는 '빵'이 상위 개념, '장미'가 하위 개념인 상속 관계를 나타낸다.
ㄷ. 계산학적 완전성에 대한 보장 없이 최대의 표현력을 활용하여 온톨로지 구축을 원하는 사용자는 OWL Lite보다는 OWL Full을 사용할 것이다.

① ㄱ ② ㄴ ③ ㄷ
④ ㄱ, ㄴ ⑤ ㄴ, ㄷ

24. 윗글과 <보기>를 바탕으로 소와의 상위 수준 온톨로지에 대해 이해한 것으로 적절하지 않은 것은?

<보 기>

소와의 상위 수준 온톨로지를 그림으로 나타내면 다음과 같다.



⊤는 세상에 존재하는 모든 것들의 집합을, ⊥는 공집합을 뜻한다. ⊤ 바로 아래 원초적 개념으로 'Independent'와 'Relative'와 'Mediating', 'Physical'과 'Abstract', 'Continuant'와 'Occurrent' 이렇게 7가지가 있다. 하나의 선으로 연결된 두 개념 중 위쪽이 상위 개념, 아래쪽이 하위 개념이다.

한편 상속 관계는 추이성(transitivity)을 갖는 대표적인 관계다. 즉 A, B, C가 각각 개념이라 할 때, 하위 개념 A가 상위 개념 B와 상속 관계를 맺고 하위 개념 B가 상위 개념 C와 상속 관계를 맺으면, 하위 개념 A는 상위 개념 C와 상속 관계를 맺는다.

① 상위 개념으로 원초적 개념을 단 한 개만 갖는 개념은 없고, 오직 2개의 원초적 개념을 갖는 개념은 모두 6개다.
② ⊤는 세상에 존재하는 모든 것들이므로 이 개념은 존재하는 모든 속성을 다 가지고 있고, ⊥에는 어떠한 개체도 속하지 않으므로 이 개념은 어떠한 속성도 갖지 않는다.
③ 'Continuant'와 'Occurrent'의 공통 하위 개념은 오직 ⊥뿐이므로, 'Continuant'의 속성과 'Occurrent'의 속성을 모두 갖는 개체는 존재하지 않는다.
④ 'Object'는 'Actuality'의 하위 개념이고 또한 'Continuant'의 하위 개념이기도 하므로, 'Actuality'의 속성과 'Continuant'의 속성을 모두 물려받는다.
⑤ 'Process'는 'Actuality'의 하위 개념이고 'Actuality'는 'Physical'의 하위 개념인데, 상속 관계는 추이성을 가지므로, 'Process'는 'Physical'의 하위 개념이다.

- 정답: 22.④ 23.③ 24.②

◆ 21 LEET 언어이해 1~3번

[1~3] 다음 글을 읽고 물음에 답하시오.

비즈니스 프로세스는 고객 가치 창출을 위해 기업 또는 조직에서 업무를 처리하는 과정을 말한다. 업무 처리 과정을 업무흐름도로 도식화하는 과정을 프로세스 모델링이라 하며, 그 결과물을 프로세스 모델이라고 한다. 프로세스 모델은 업무 처리 활동 및 활동들 간의 경로로 구성된다. 프로세스 모델이 효율적으로 작동하고 있는지를 확인, 분석, 수정·보완, 개선하는 작업이 필요한데, 프로세스 마이닝은 그중 한 기법이다. 프로세스 마이닝은, 시뮬레이션처럼 실제 이벤트 로그 수집 이전에 정립한 프로세스 모델 중심 분석기법과, 데이터 마이닝처럼 프로세스를 고려하지 않는 데이터 중심 분석기법을 연결하는 역할을 한다.

프로세스 마이닝은 정보시스템을 통해 확보한 이벤트 로그에서 프로세스에 관련된 가치 있는 정보를 추출하는 것이다. 이벤트 로그란 정보시스템에 축적된 비즈니스 프로세스 수행 기록인데, 이것이 프로세스 마이닝의 출발점이 된다. 이벤트 로그는 행과 열로 표현되는 이차원 표 형태이다. 업무 활동으로 발생한 이벤트는 행으로 추가되며, 각 열에는 이벤트의 속성들이 기록된다. 이때 기록되는 속성으로 필수적인 것은 사례 ID, 활동명, 발생 시점이며, 다양한 분석을 위해 그 외 속성들도 추가될 수 있다. 이벤트 로그는 사용자에게 도움이 되는 정보를 직접 제공할 수 없는 원데이터이므로, 그것을 우리가 사용할 수 있는 정보로 변환해 주어야 한다. 프로세스 마이닝에는 프로세스 발견, 적합성 검증, 프로세스 향상의 세 가지 유형이 있다.

프로세스 발견이란 프로세스 분석가가 알고리즘을 통해 이벤트 로그로부터 프로세스 모델을 도출하는 것을 말하는데, 이때 분석가는 별다른 업무 지식 없이도 작업을 수행할 수 있다. 만일 도출된 프로세스 모델이 복잡하여 유의미한 분석이 곤란할 경우, 퍼지 마이닝이나 클러스터링 기법을 활용할 수 있다. 퍼지 마이닝은 실행 빈도가 낮은 활동을 제거 또는 병합하거나, 그 활동들 간의 경로를 제거함으로써 프로세스 모델을 단순화해 주는 기법이다. 이때 프로세스 모델에 나타난 활동과 경로에 대한 임곗값을 설정하여 모델의 복잡도를 조절할 수 있다. 클러스터링은 특성이 유사한 사례들을 같은 그룹으로 묶어주는 기법이다. 전체 이벤트 로그를 대상으로 프로세스를 도출할 때 복잡한 프로세스 모델이 도출될 경우, 이 기법을 적용하여 이벤트 로그를 여러 개로 나눌 수 있다. 이렇게 세분화된 이벤트 로그에 프로세스 발견 기법을 적용하면, 프로세스 모델의 복잡도가 줄어든다.

적합성 검증이란 기존의 프로세스 모델과 이벤트 로그 분석에서 도출된 결과를 비교하여 어느 정도 일치하는지를 확인하는 것이다. 이때 기존의 프로세스 모델과 이벤트 로그에서 도출된 결과물이 불일치하는 경우가 발생하는데, 먼저 기존의 프로세스 모델이 적절함에도 불구하고 업무 담당자가 이를 준수하지 않는 경우를 들 수 있다. 이 경우에는 현실 세계의 실제 업무 수행 실태를 교정해야 한다. 이와 달리 이벤트 로그의 분석 결과물이 더 적절한 것으로 판단되는 경우에는 기존의 프로세스 모델을 수정할 필요가 있다.

프로세스 향상에는 두 유형이 있다. 하나는 기존의 프로세스 모델을 '수정'하는 것이며, 다른 하나는 업무 수행 시간 및 담당자 등 이벤트 로그 분석에서 얻은 부가적 정보를 추가하여 발견된 프로세스 모델을 '확장'하는 것이다. 확장의 예로는 이벤트 로그로부터 도출된 프로세스 모델에 프로세스 내 병목지점과 재작업 흐름을 시각화하는 것을 들 수 있다.

프로세스 마이닝은 데이터 과학에 근거를 두고 프로세스 분석가가 업무 전문가와 협업하여 기업이 수행하는 비즈니스 프로세스에 대한 문제점을 진단하고 개선 방안을 도출하는 데 기여할 수 있다.

1. 윗글과 일치하는 것은?

① 이벤트 로그는 프로세스 마이닝의 출발점이지만 그 자체로는 유용한 정보라 할 수 없다.

② 업무 전문가의 충분한 지식 없이 이벤트 로그로부터 프로세스 모델을 도출하기는 어렵다.

③ 프로세스 발견은 프로세스에 내재된 업무 관련 규정을 이벤트 로그로부터 도출하는 것이다.

④ 클러스터링은 복잡한 프로세스 모델을 여러 개의 세부 프로세스 모델로 구분해 주는 기법이다.

⑤ 이벤트 로그에서 업무 담당자를 파악하여 기존의 프로세스 모델에 활동과 경로를 추가하는 것은 프로세스 수정이다.

2. '프로세스 마이닝'에 대해 추론한 것으로 적절하지 않은 것은?

① 프로세스 마이닝을 도입하면 내부 규정의 준수 여부에 대한 감독이 용이해진다.

② 프로세스 마이닝을 통해 기존의 프로세스 모델이 실제로 어떻게 수행되는가를 파악할 수 있다.

③ 프로세스 마이닝은 판에 박힌 단순한 업무뿐 아니라 비정형적인 업무 처리 과정의 분석에도 활용된다.

④ 프로세스 마이닝은 예상된 이벤트 로그에 적용할 프로세스 모델 중심의 업무 성과 분석 및 개선 기법이다.

⑤ 프로세스 마이닝은 기존의 프로세스 모델뿐 아니라 발견으로 도출된 프로세스 모델을 향상하는 데에도 활용된다.

3. <보기>의 사례에 프로세스 마이닝을 적용할 때 가장 적절한 것은?

<보 기>

○○병원에서는 외래 환자의 과도한 대기 시간을 줄이고 의료 서비스의 품질을 개선하기 위해 외래 환자 진료 프로세스를 분석하고자 한다. 이 병원에서는 질환별로 진행해야 하는 표준 진료 프로세스를 임상진료 지침으로 수립해 두고 있다. 프로세스 마이닝 도구를 사용하여 프로세스 모델을 도출하였더니 지나치게 복잡한 프로세스 모델이 도출되어 분석이 곤란한 상황이다. 또한 환자의 민감한 개인 의료정보가 저장된 이벤트 로그를 프로세스 분석가에게 제공할 경우 정보 보호 및 프라이버시 이슈가 존재하고, 병원의 기밀이 유출될 우려가 제기되어 이를 해결하고자 한다.

① 복잡도 문제를 해결하기 위해 연령 및 질환을 기준으로 이벤트 로그의 사례를 클러스터링 하려면 필수적 속성만 이벤트 로그에 있어도 된다.

② 적합성 검증 결과 기존의 프로세스 모델과 이벤트 로그 분석 결과가 불일치하면 의료진에 대한 제재 조치나 지침 재교육이 필수적이다.

③ 이벤트 속성의 임곗값을 조절하여 빈번하게 수행되는 진료 프로세스 수행 패턴을 파악할 수 있다.

④ 환자의 개인정보 보호를 위해 사례 ID를 제외하고 이벤트 로그를 작성해야 한다.

⑤ 외래 환자의 대기 시간 분석을 위해서는 프로세스 확장이 필요하다.

- 정답: 1.① 2.④ 3.⑤

[16~18] 다음 글을 읽고 물음에 답하시오.

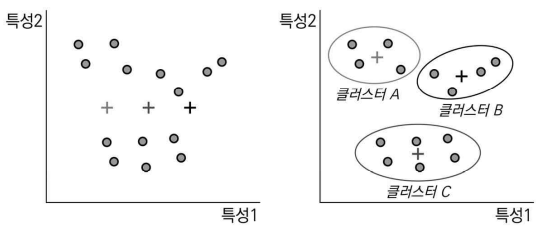
대규모 데이터를 분석하여 데이터 속에 숨어 있는 유용한 패턴을 찾아내기 위해 다양한 기계학습 기법이 활용되고 있다. 기계학습을 위한 입력 자료를 데이터 세트라고 하며, 이를 분석하여 유용하고 가치 있는 정보를 추출할 수 있다. 데이터 세트의 각 행에는 개체에 대한 구체적인 정보가 저장되며, 각 열에는 개체의 특성이 기록된다. 개체의 특성은 범주형과 수치형으로 구분되는데, 예를 들어 '성별'은 범주형이며, '체중'은 수치형이다.

기계학습 기법의 하나인 클러스터링은 데이터의 특성에 따라 유사한 개체들을 묶는 기법이다. 클러스터링은 분할법과 계층법으로 나뉘는데, 이 둘은 모두 거리 개념에 기초하고 있다. 가장 많이 사용되는 거리 개념은 기하학적 거리이며, 두 개체 사이의 거리는 n차원으로 표현된 공간에서 두 개체를 점으로 표시할 때 두 점 사이의 직선거리이다. 거리를 계산할 때 특성들의 단위가 서로 다른 경우가 많은데, 이런 경우 특성 값을 정규화할 필요가 있다. 예를 들어 특정 과목의 학점과 출석 횟수를 기준으로 학생들을 묶을 경우 두 특성의 단위가 다르므로 두 특성 값을 모두 0과 1 사이의 값으로 정규화하여 클러스터링을 수행한다. 또한 범주형 특성에 거리 개념을 적용하려면 이를 수치형 특성으로 변환해야 한다.

분할법은 전체 데이터 개체를 사전에 정한 개수의 클러스터로 구분하는 기법으로, 모든 개체는 생성된 클러스터 가운데 어느 하나에 속한다. <그림 1>에서 (b)는 (a)에 제시된 개체들을 분할법을 통해 세 개의 클러스터로 묶은 예이다. 분할법에서는 클러스터에 속한 개체들의 좌표 평균을 계산하여 클러스터 중심점을 구한다. 고전적인 분할법인 K-민즈 클러스터링(K-means clustering)에서는 거리 개념과 중심점에 기반하여 다음과 같은 과정으로 알고리즘이 진행된다.

1) 사전에 K개로 정한 클러스터 중심점을 임의의 위치에 배치하여 초기화한다.
2) 각 개체에 대해 K개의 중심점과의 거리를 계산한 후 가장 가까운 중심점에 해당 개체를 배정하여 클러스터를 구성한다.
3) 클러스터 별로 그에 속한 개체들의 좌표 평균을 계산하여 클러스터의 중심점을 다시 구한다.
4) 2)와 3)의 과정을 반복해서 수행하여 더 이상 변화가 없는 상태에 도달하면 알고리즘이 종료된다.

분할법에서는 이와 같이 개체와 중심점과의 거리를 계산하여 클러스터에 개체를 배정하므로 두 개체가 인접해 있더라도 가장 가까운 중심점이 서로 다르면 두 개체는 상이한 클러스터에 배정된다.

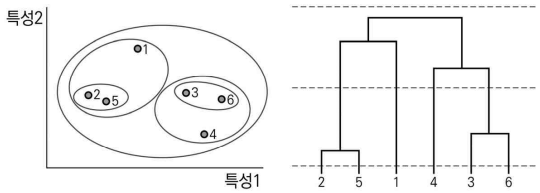


(a) 중심점 초기화 (b) 분할법 최종 결과

<그림 1> 분할법의 예

클러스터링이 잘 수행되었는지 확인하려면 클러스터링 결과를 평가하는 품질 지표가 필요하다. K-민즈 클러스터링의 경우 품질 지표는 개체와 그 개체가 해당하는 클러스터의 중심점 간 거리의 평균이다. K-민즈 클러스터링에서 K가 정해졌을 때 개체와 해당 중심점 간 거리의 평균을 최소화하는 '전체 최적해'는 확정적으로 보장되지 않는다. 알고리즘의 첫 번째 단계인 초기화를 어떻게 하느냐에 따라 클러스터링 결과가 달라질 수 있으며, 경우에 따라 좋은 결과를 찾는 데 실패할 수도 있다. 따라서 전체 최적해를 얻을 확률을 높이기 위해, 서로 다른 초기화를 시작으로 클러스터링 알고리즘을 여러 번 수행하여 나온 결과 중에 좋은 해를 찾는 방법이 흔히 사용된다. 그런데 K-민즈 클러스터링 알고리즘의 한 가지 문제는 클러스터의 개수인 K를 미리 정해야 한다는 것이다. K가 커질수록 각 개체와 해당 중심점 간 거리의 평균은 감소한다. 극단적으로 모든 개체를 클러스터로 구분할 경우 개체가 곧 중심점이므로 이들 사이의 거리의 평균값은 0으로 최소화되지만, 클러스터링의 목적에 부합하는 유용한 결과라고 보기 어렵다. 따라서 작은 수의 K로 알고리즘을 시작하여 클러스터링 결과를 구한 다음 K를 점차 증가시키면서 유의미한 품질 향상이 있는지 확인하는 방법이 자주 사용된다.

한편, 계층법은 클러스터 개수를 사전에 정하지 않아도 되는 장점이 있다. <그림 2>와 같이 개체들을 거리가 가까운 것들부터 차근차근 집단으로 묶어서 모든 개체가 하나로 묶일 때까지 추상화 수준을 높여가는 상향식으로 알고리즘이 진행되어 계통도를 산출한다. 따라서 계층법은 개체들 간에 위계 관계가 있는 경우에 효과적으로 적용될 수 있다. 계통도에서 점선으로 표시된 수평선을 아래위로 이동해 가면서 클러스터링의 추상화 수준을 변경할 수 있다.



(a) 상향식 그룹화 (b) 계통도

<그림 2> 계층법의 예

16. 윗글의 내용과 일치하는 것은?

① 클러스터링은 개체들을 묶어서 한 개의 클러스터로 생성하는 기법이다.

② 분할법에서는 클러스터링 수행자가 정확한 계산을 통해 초기 중심점을 찾아낸다.

③ 분할법은 하향식 클러스터링 기법이므로 한 개체가 여러 클러스터에 속할 수 있다.

④ 계층법으로 계통도를 산출할 때 클러스터 개수는 미리 정하지 않는다.

⑤ 계층법의 계통도에서 수평선을 아래로 내릴 경우 추상화 수준이 높아진다.

17. K-민즈 클러스터링에 대해 추론한 것으로 적절하지 않은 것은?

① 특성이 유사한 두 개체가 서로 다른 클러스터에 배치될 수 있다.

② 초기 중심점의 배치 위치에 따라 클러스터링의 품질이 달라질 수 있다.

③ 클러스터 개수를 감소시키면 클러스터링 결과의 품질 지표 값은 증가한다.

④ 초기화를 다르게 하면서 알고리즘을 여러 번 수행하면 전체 최적해가 결정된다.

⑤ K를 정하여 알고리즘을 진행하면 각 클러스터의 중심점은 결국 고정된 점에 도달한다.

18. <보기>의 사례에 클러스터링을 적용할 때 적절하지 않은 것은?

<보 기>

○○기업에서는 표적 시장을 선정하여 마케팅을 실행하기 위해 전체 시장을 세분화하고자 한다. 시장 세분화를 위해 특성이 유사한 고객을 묶는 기계학습 기법 도입을 검토 중이다. 이 기업에서는 고객의 거주지, 성별, 나이, 소득 수준 등 인구통계학적인 정보와 라이프 스타일에 관한 정보 등을 보유하고 있다.

① 고객 정보에는 수치형이 아닌 것도 있어 특성의 유형 변환이 요구된다.

② 고객 특성은 세분화 과정을 통해 계통도로 표현 가능하므로 계층법이 효과적이다.

③ K-민즈 클러스터링 알고리즘을 실행하려면 세분화할 시장의 개수를 먼저 정해야 한다.

④ 나이와 소득수준과 같이 단위가 다른 특성을 기준으로 시장을 세분화할 경우 정규화가 필요하다.

⑤ 모든 고객을 별도의 세분화된 시장들로 구분하여 1:1 마케팅을 할 경우 K-민즈 클러스터링의 품질 지표 값은 0이다.

- 정답: 16.④ 17.④ 18.②