[4~6] 언어이해 22` 16~18

4	5	6
④	④	②

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.

- 데이터 세트: 말 그대로 데이터(자료). 기계학습을 위한 자료. 이것을 분석해 유용한 정보를 얻는다.
- 행: 개체에 대한 구체적인 정보
- 열: 개체의 특성
- 개체의 특성: 범주형과 수치형으로 구분된다.

□ 개체의 특성 중 '체중'처럼 수치로 나타내는 것은 수치형이고, '성별'처럼 개체가 속하는 범주를 나타내는 것은 범주형이다.

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 특성값을 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.

- 클러스터링: 특성이 유사한 개체들을 묶는 기계학습 기법
- 분할법과 계층법으로 나뉘고 거리 개념에 기초한다.
- 거리 개념 중 기하학적 거리가 가장 많이 사용된다.
- 두 개체 사이의 거리: n차원 공간에서 개체를 점으로 표현할 때 두 점 사이의 직선 거리
- 수치의 단위가 다른 경우 정규화해야 한다.
- 정규화: 0과 1 사이의 값으로 바꾸는 과정
- 범주형 특성은 수치형 특성으로 변환해야 한다.

□ 클러스터링은 특성이 유사한 개체들을 거리 개념에 기초에 묶는 기법이다.

□ 거리 개념 적용을 위해 정규화 과정 또는 범주형 특성의 수치형 특성으로의 변환 과정을 거칠 수 있다.

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.

- 분할법: 사전의 정한 개수의 클러스터로 구분
- 사전에 클러스터 개수를 정해준다.
- 클러스터에 속하지 않는 개체는 없다.
- K-민즈 클러스터링(분할법)
- 임의의 중심점 생성(= 초기화) → 가장 가까운 중심점에 개체 배정 → 클러스터에 속한 개체들의 좌표 평균을 새로운 중심점으로 설정 → 중심점이 고정될 때까지 반복

□ K-민즈 클러스터링

- 클러스터 중심점 개수 = 클러스터의 개수
- 두 개체가 인접해 있더라도(= 특성이 유사해도) 다른 클러스터에 배정될 가능성이 있다.

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.

- 품질 지표: 클러스터링 결과를 평가하는 지표
- K-민즈 클러스터링: 개체와 클러스터의 중심점 간 거리의 평균

- 전체 최적해: 개체와 중심점 간 거리의 평균(= 품질 지표)이 최소인 클러스터링 결과. 그렇다면 이것은 가능한 최상의(best) 결과인가?

□ K-민즈 클러스터링에서는 품질 지표가 낮을수록 전체 최적해에 가깝다.

□ K-민즈 클러스터링에서 전체 최적해는 확정적으로 보장되지 않는다. K-민즈 클러스터링을 수행하면 항상 K개의 클러스터로 클러스터링 되지만, 이것의 수행으로 항상 전체 최적해를 구할 수 있는 것은 아니다. K-민즈 클러스터링이 보장하는 것과 보장하지 않는 것을 구분해서 이해하자.

□ 클러스터링 결과는 초기에 임의로 정한 중심점의 위치와 K 값에 따라 달라진다.

❖ 따라서 좋은 결과를 얻기 위해
1. 서로 다른 초기화를 시작으로 여러 번 반복해 좋은 해를 찾는다.
2. 서로 다른 K 값으로 여러 번 반복해 좋은 해를 찾는다.

□ K가 커질수록 결과의 품질 지표는 그것의 품질 향상과 무관하게 항상 작아진다.

❖ 여기서 알 수 있는 것
1. K 값의 경우 작은 수로 시작해 점차 증가시키면서 클러스터링을 반복해 결과의 품질 향상이 있는지 확인한다.
2. 품질 지표가 작을수록 품질이 좋은 것은 아니다.

→ 전체 최적해가 항상 "가능한 최상의 결과"인 것은 아니다.

→ 품질 지표는 클러스터링 결과를 평가하는 데 쓰일 뿐 클러스터링 결과의 품질을 완벽히 반영하지 못한다.

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.

- 계층법:

- 클러스터 개수를 사전에 정하지 않는다.
- (그림 참고) 거리가 가까운 개체들을 묶고, 그것들끼리 또 다시 묶는 과정을 모든 개체가 하나로 묶일 때까지 반복한다.
- 과정이 진행될수록 묶음의 개수가 작아지고, 추상화 수준이 높아진다.
- 개체들 간에 위계 관계가 있는 경우에 효과적이다.

□ 그렇다면 클러스터의 개수는 어떻게 결정되는가?

- 클러스터링의 추상화 수준을 변경하면 이에 맞춰 클러스터의 개수가 정해진다.
- 특정 추상화 수준에서의 개체들의 묶음들이 각각 클러스터가 되는 것이다.

❖ 분할법의 경우 일단 실행되면 클러스터의 개수를 변경할 수 없지만, 계층법은 실행 후 추상화 수준을 변경해 클러스터의 개수를 조절할 수 있을 것 같다(본문에서는 점선으로 표시된 수평선을 조절한다고 표현했다).

4. 윗글의 내용과 일치하는 것은?

① 클러스터링은 개체들을 묶어서 한 개의 클러스터로 생성하는 기법이다.

클러스터링은 클러스터로 개체들을 묶는 기법이다. 이때 클러스터의 개수가 꼭 한 개일 필요는 없다. 본문에 제시된 분할법의 예시에서도 3개의 클러스터가 확인된다.

② 분할법에서는 클러스터링 수행자가 정확한 계산을 통해 초기 중심점을 찾아낸다.

초기 중심점은 임의로 결정된다. 임의로 결정된다는 것은 계산에 의한 것이라고 볼 수 없다.

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

그림1을 확인하면, 분할법에서는 한 개체가 하나의 클러스터에 속한다는 것을 쉽게 알 수 있다.

한편, 본문에서 '상향식'이라는 속성은 계층법에 관한 설명에서 확인할 수 있다. 계층법은 시작 시점에 낮았던 추상화 수준이 과정을 거듭할수록 높아지는 상향식 클러스터링 기법이다. 이때 추상화 수준은 묶음의 개수, 즉 클러스터의 개수를 결정한다. 추상화 수준의 특정 범위에 클러스터의 개수가 일대일 대응되는 것이다(그림2 참고).
분할법에서는 클러스터의 개수를 사전에 정해둔다. 이는 분할법에서는 추상화 수준이 클러스터링 전에 정해진다는 뜻이다. 따라서 클러스터링 과정이 진행되며 추상화 수준이 높아지지도, 낮아지지도 않으므로 분할법은 상향식과 하향식 클러스터링 기법 중 어느 것도 아니다.

※ 밑줄에 대한 추가 설명:
예를 들어 (계층법에서) 추상화 수준을 0부터 10까지 산정한다고 할 때, 0 이상 3 미만에서는 클러스터가 3개, 3 이상 8 미만에서는 클러스터가 2개, 8 이상 10 이하에서는 클러스터가 1개일 수 있다.

④ 계층법으로 계통도를 산출할 때 클러스터 개수는 미리 정하지 않는다.

계층법에서는 사전에 클러스터 개수를 정하지 않고 모든 개체가 하나로 묶일 때까지 계통도를 그린다. 따라서 계통도를 산출할 때에도 클러스터의 개수는 정해지지 않으므로 옳은 선지다.

⑤ 계층법의 계통도에서 수평선을 아래로 내릴 경우 추상화 수준이 높아진다.

계층법은 개체들을 묶은 묶음의 개수가 점점 작아지는 방향으로 진행된다. 따라서 그림2의 계통도에서 묶음의 개수가 많은 아래쪽이 시작 시점, 모든 개체가 하나로 묶인 위쪽이 종료 시점일 것이다. 개체들을 묶는 작업이 거듭될수록 추상화 수준은 높아진다고 했으므로 시작 시점에서 종료 시점으로 갈수록, 즉 계통도에서 위쪽으로 갈수록 추상화 수준은 높아진다. 따라서 수평선을 아래로 내릴 경우 추상화 수준은 낮아진다.

5. K-민즈 클러스터링에 대해 추론한 것으로 적절하지 않은 것은?

① 특성이 유사한 두 개체가 서로 다른 클러스터에 배치될 수 있다.

본문에 따르면 인접한 두 개체가 서로 다른 클러스터에 배치될 수 있다. 그림 1의 (a)를 참고할 때, 두 개체에 대응하는 두 점의 거리가 서로 가깝다는 것은 특성이 유사하다는 의미이므로 적절하다.

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

본문에서는 알고리즘의 첫 단계인 초기화를 어떻게 하느냐에 따라 클러스터링 품질이 달라질 수 있다고 했다. 초기화 단계에서는 각 클러스터의 중심점을 임의로 배치하므로, 이것을 어떻게 하느냐는 곧 초기 중심점을 어떻게 배치하느냐를 의미한다. 따라서 적절하다.

③ 클러스터 개수를 감소시키면 클러스터링 결과의 품질 지표 값은 증가한다.

본문에 따르면 '클러스터의 개수 = K 값'이고 K 값이 증가하면 품질 지표 값은 항상 감소한다. 따라서 이것의 반대인 K 값이 감소할 때에는 품질 지표 값이 증가할 것이라는 추론은 적절하다.

※ 무언가를 느끼는 학생들을 위해:
그대들이 생각하는 바와 같이 K 값이 감소할 때 품질 지표 값이 감소하는 경우가 존재할 수 있다. 따라서 해당 선지는 사실과 **일치**하지는 않는다. 그러나 본문에 제시된 사실 관계를 뒤집은 것으로 여전히 유효한 **추론**이다.

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

K-민즈 클러스터링에서 전체 최적해는 보장되지 않는다. 따라서 초기화를 다르게 해서 알고리즘을 아무리 많이 수행한다고 하더라도 전체 최적해가 결정된다고 볼 수는 없다.

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

각 클러스터의 중심점이 고정된 점에 도달한다는 것은 어느 시점에서는 알고리즘의 각 과정을 반복해도 중심점의 위치가 변하지 않는다는 것을 의미한다. 알고리즘의 구조상 알고리즘을 무한히 시행했을 때에도 중심점의 위치가 고정되지 않는 경우는 존재할 수 없으므로 적절한 추론이다.

6. 〈보 기〉의 사례에 클러스터링을 적용할 때 적절하지 <u>않은</u> 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.

클러스터링은 거리 개념에 의존하므로 클러스터링을 적용하려면 범주형을 수치형으로 변환해야 한다. 성별은 범주형 정보이므로, 고객 정보에는 수치형이 아닌 것이 있으므로 이것에 클러스터링을 적용하려면 특성의 유형 변환이 이루어져야 한다.

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

본문에 따르면 개체들 간에 위계 관계가 있을 때 계층법은 효과적이다. 따라서 고객 특성을 계통도로 표현 가능한 것은 맞지만, 이것만으로 계층법이 효과적이라고 판단할 수는 없다.

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

보기의 기업은 전체 시장을 세분화하는 목적을 갖고 유사한 특성을 가진 고객을 묶는 클러스터링을 했으므로 각 클러스터가 곧 특정 시장이 될 것이다. K-민즈 클러스터링 알고리즘은 사전에 클러스터의 개수를 정해야 하므로 보기의 사례에서는 시장의 개수를 정해야 한다.

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

본문에서 단위가 다른 특성의 경우 0과 1 사이의 값으로 정규화 하는 과정이 필요하다고 했으므로 적절하다.

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

모든 고객을 별도의 세분화된 시장들로 구분해 1:1 마케팅을 하겠다는 것은 클러스터의 개수가 고객의 개수와 일치한다는 것을 의미한다. 즉 고객 한명 한명이 각각 하나의 클러스터를 이루는 것으로 고객(= 개체)이 클러스터의 중심점이 된다. 따라서 개체와 클러스터의 중심점 간 거리의 평균인 품질 지표 값은 0이 될 것이다.