★ LEET연계(2) - 수특 170 ★

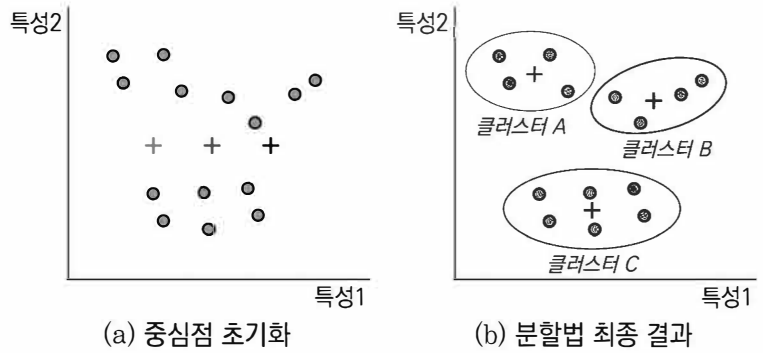
대규모 데이터를 분석하여 데이터 속에 숨어 있는 유용한 패턴을 찾아내기 위해 다양한 기계학습 기법이 활용되고 있다. 기계학습을 위한 입력 자료를 데이터 세트라고 하며, 이를 분석하여 유용하고 가치 있는 정보를 추출할 수 있다. 데이터 세트의 각 행에는 개체에 대한 구체적인 정보가 저장되며, 각 열에는 개체의 특성이 기록된다. 개체의 특성은 범주형과 수치형으로 구분되는데, 예를 들어 '성별'은 범주형이며, '체중'은 수치형이다.

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.

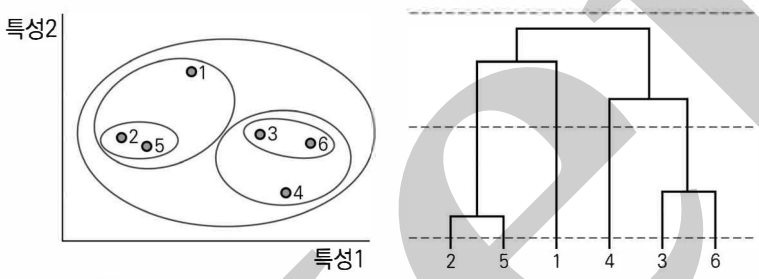


(a) 중심점 초기화 (b) 분할법 최종 결과

<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.