**클러스터링(Clustring)**

**1. 클러스터 탐지(군집분석)의 정의**

가. 전체 데이터를 몇 개의 집단으로 그룹화하여 각 집단의 성격을 파악

ㄴ 데이터 전체의 구조에 대한 이해를 돕고자 하는 분석법

나. 모집단 또는 범주에 대한 사전 정보가 없는 경우

ㄴ 관측 값들 사이의 거리 또는 유사성을 이용하는 분석법

다. 주어진 데이터들의 특성을 고려해 데이터 집단(클러스터)을 정의하고 데이터 집단의 대표할 수 있는 대표점을 찾는 것으로 데이터 마이닝의 한 방법

**2. 군집화**

가. N개의 관찰치를 대상으로 p개의 변수를 측정했을 때 관측한 p개의 변수 값을 이용하여

N개의 관찰치 사이의 유사성 또는 비 유사성의 정도를 측정하여 가까운 순서대로 군집화.

|  |  |
| --- | --- |
| **구분** | **설명** |
| 1. 군집화의 기준 | 1.1 동일한 군집에 속하는 개체는 여러 속성이 비슷하고  1.2 서로 다른 군집에 속한 관찰치는 그렇지 않도록 구성 |
| 2. 군집화를 위한 변수, | 2.1 인구 통계적 변수: 성별, 나이, 거주지, 직업, 소득, 교육, 종교 등  2.2 구매 패턴 변수: 상품, 주기, 거래액 등 |

나. 군집분석의 활용

- 군집분석은 속성이 비슷한 잠재 고객들끼리 그룹화 하여 시장을 세분화 하는 방법에 자주 활용

|  |  |  |
| --- | --- | --- |
| http://www.jidum.com/upload/ckeditor/2018/01/20180123132512867.png | | |
| **구분** | **설명** | |
| 기업의 수익에 기여 정도 | - 우수고객의 인구통계적 요인, 생활패턴 파악  - 개별고객에 대한 맞춤관리 | |
| 구매패턴 | - 신상품 판촉, - 교차판매를 위한 상품 구성 | |

다. 군집분석의 특징

|  |  |
| --- | --- |
| **구분** | **설명** |
| 탐색적인 기법 | 주어진 자료에 대한 사전정보 없이 의미 있는 자료구조를 찾아 낼 수 있음 |
| 다양한 형태의  데이터에 적용가능 | 거리만 잘 정의되면 모든 종류의 자료에 적용 할 수 있음 |
| 분석자에 의존 | 자료의 사전 정보 없이 자료를 파악하는 방법으로, 분석자의 주관에 결과가 달라짐 |
| 분석방법 용이 | 분석방법의 적용이 쉬움 |
| 가중치 및 거리 | 가중치와 거리의 정의가 어려움 |
| 결과 해석 | 결과의 해석이 어려움 |
| 초기 군집수 K 결정 | 초기 군집수 K의 결정이 어려움 |

**3. 군집분석 알고리즘**

가. 비계층적 군집분석

(1) 관측 값들을 몇 개의 군집으로 나누기 위하여 주어진 판정기준을 최적화하는 방식

(2) 대표적으로 K-means 알고리즘

나. K-means 알고리즘

(1) 사전에 결정된 군집수 K에 기초

(2) 전체 데이터를 상대적으로 유사한 K개의 군집으로 분할

(3) 군집수 K는 반복적으로 K를 달리 사용해봐서 가장 적합한 K를 결정

(4) k-평균 군집법은 계층적 군집법에 비하여 계산량이 적음

(5) 대용량 데이터를 빠르게 처리할 수 있음

다. K-means 절차

(1) 군집의 수 K를 정의 및 초기 K개 군집의 중심 랜덤 선택

(2) 각 관측 값들을 가장 가까운 중심의 군집에 할당

(3) 새로운 군집의 중심 계산

(4) 재정의된 중심값 기준으로 다시 거리기반 군집 재분류

라. K-means 알고리즘 장단점

|  |  |  |
| --- | --- | --- |
| **구분** | **세부사항** | **설명** |
| 장점 | 탐색적인 기법 | 군집분석은 그 자체가 대용량 데이터에 대한 탐색적인 기법 주어진 데이터의 내부구조에 대한 사전적인 정보 없이 의미 있는 자료구조를 찾아낼 수 있는 방법 |
| 다양한 형태의 데이터에 적용 가능 | 분석을 위해서는 기본적으로 관찰치 간의 거리를 데이터형태에 맞게만 정의하면, 거의 모든 형태의 데이터에 대하여 적용이 가능한 방법 |
| 분석방법의 적용 용이성 | 대부분의 군집방법이 분석대상 데이터에 대해 사전정보를 거의 요구하지 않음 적용 유리  사전에 특정 변수에 대한 역할 정의가 필요하지 않고 관찰치 사이의 거리만이 분석에 필요한 입력자료로 사용. |
| 단점 | 초기 군집수의 결정 | k-평균 군집분석에서는 사전에 정의된 군집 수를 기준으로 동일한 수의 군집을 찾게 되므로 만일 군집수 k가 원 데이터구조에 적합하지 않으면 좋은 결과를 얻을 수 없다. |
| 결과해석의 어려움 | 탐색적인 분석방법으로의 장점을 가지고 있는 반면에 사전에 주어진 목적이 없으므로 결과를 해석하는데 있어서 어려움 존재 |

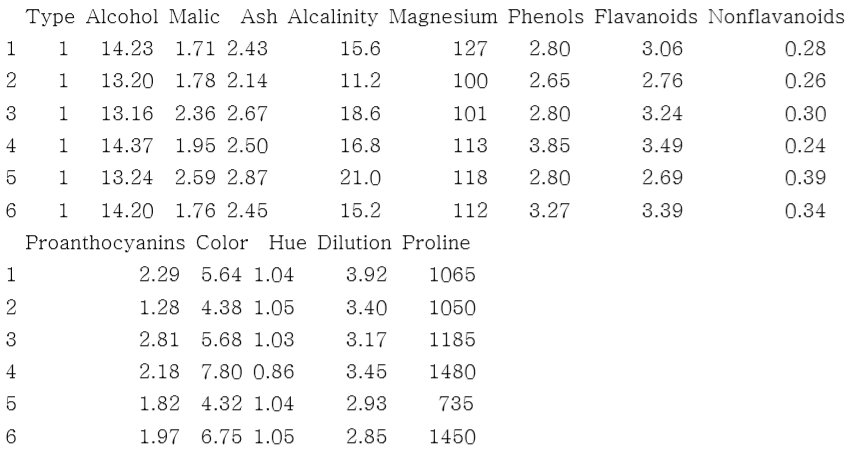
**마. 군집분석(K-means)의 활용 방안**

(1) 시장과 고객 분석, 패턴인식, 공간데이터 분석, Text Mining 등

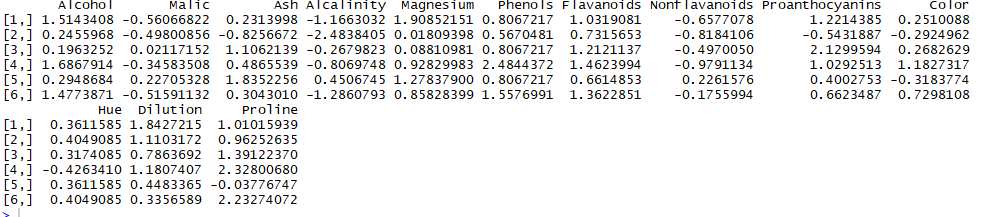
(2) 최근에는 패턴인식, 음성인식의 기본 알고리즘으로 활용

**4. K-MEANS 알고리즘 실제 데이터**

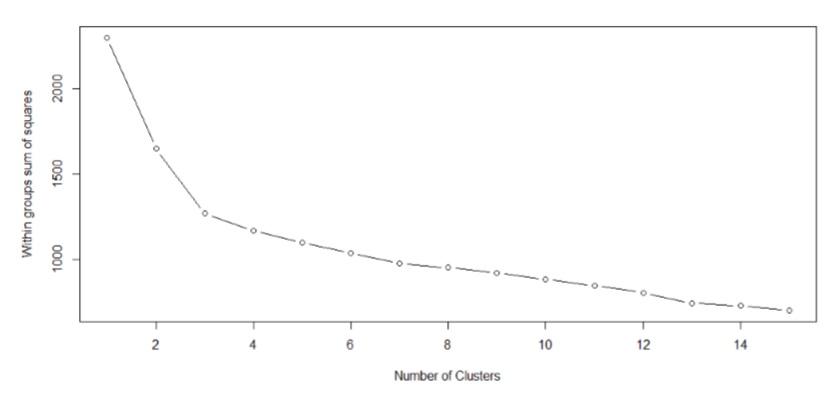
가. 178개 이탈리안 와인에 대한 13가지의 화학적 성분을 측정한 데이터



1. 변수의 측정 단위가 각각 다르므로 군집분석을 수행하기 전에 scale() 함수 사용 표준화 진행

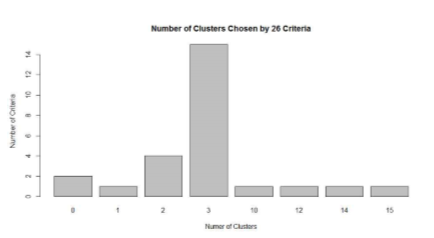


1. 군집수를 정하기 위해 wssplot() 제곱합의 그래프를 그려 군집수를 결정

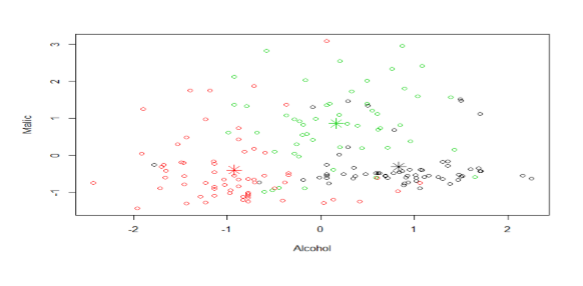


1. 최적의 군집수를 정하기 위해 NBCLUST() 함수를 이용하여 한번더 군집수 체크

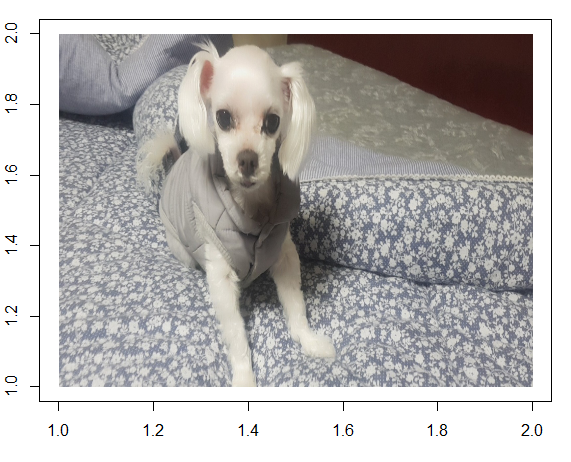
가운데 약 15개의 지수가 을 최적의 군집수로 투표한 결과를 보여줌



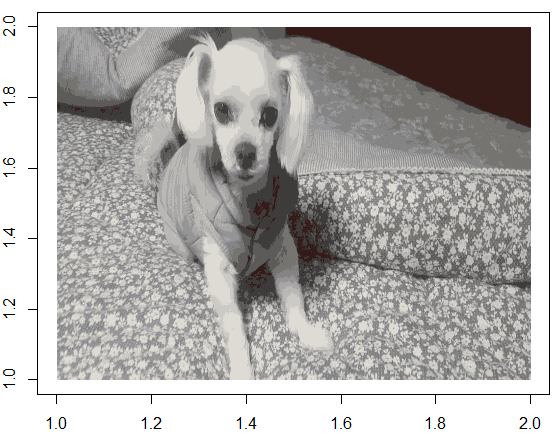
1. 군집의 수를 3으로하여 K-means를 수행한 결과



나. 이미지 클러스터링 – 이미지 분할 작업



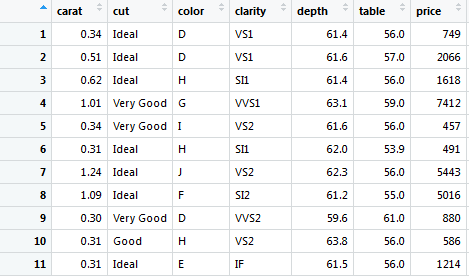
1. 비슷한 색상들을 대표 색상으로 그룹화하는 작업을 진행, 이미지의 각 픽셀을 하나의 데이터 포인트로 간주
2. 각각의 색상의 범위를 정규화하여 [0,1]사이의 값만 가지도록 설정
3. K-means 알고리즘을 적용하여 픽셀데이터를 R,G,B로 표현되는 중심 벡터로 표현



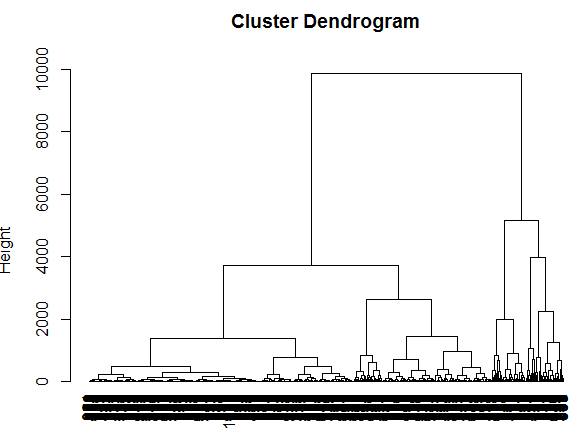
1. 위의 그림은 이미지에 K-means를 적용한 결과
2. 각각의 점들을 얻어진 중심 픽셀의 대표 색으로 변경한 뒤에 다시 그렸을 때의 결과

**5. 계층적 군집분석**

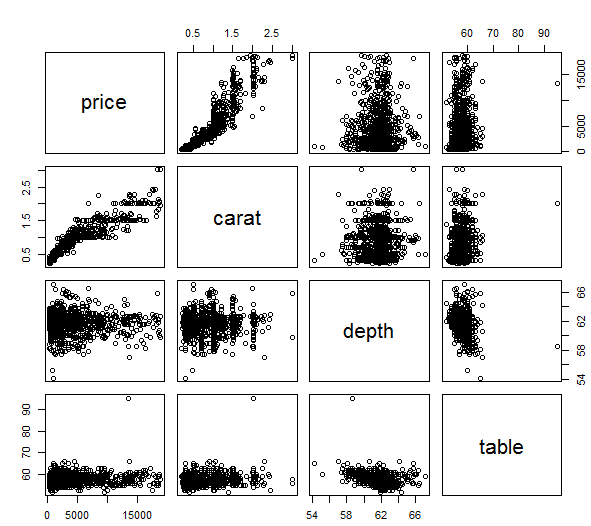
가. 다이아몬드 데이터(캐럿, 색깔, 컷, 테이블, 가격, 뎁스(높이), 선명도)



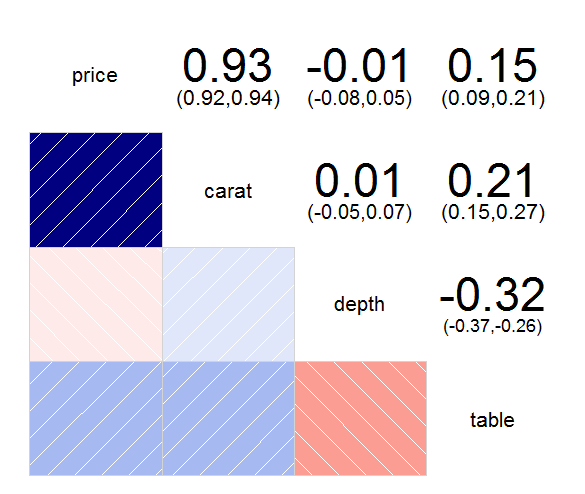
1. 군집 파악용 변수만 일부 추출 (price, carat, depth, table)
2. 계층적 클러스터 방식으로는 평균연결법 사용



1. 위의 덴드로그램을 바탕으로 컷포인트를 6000정도로보고 군집을 3개로 설정
2. 위의 데이터를 비계층적 군집분석인 k-means를 이용하여 3개의 그룹으로 군집화
3. 3개의 그룹으로 나누어진 변수들간의 상관관계 파악을 위해 피어슨 상관계수 사용



1. 변수들간의 상관관계를 수치로 쉽게 파악하기 위해 corrgram함수를 사용한다



1. 상관관계 표를 보았을 때 가격과 캐럿이 상관관계가 제일 높다는 것을 알 수 있다
2. 캐럿과 가격의 군집을 시각화 했을 때 3개의 군집이 나누어진 것을 확인 가능

