## Algorytm k – średnich

Algorytm k-średnich podziału zbioru obiektów bazujący na średniej wartości (środka ciężkości) w skupieniu.

Wejście: Liczba skupień „k”, liczba obiektów „n”.

Wyjście: Zbiór „k” skupień minimalizujący kryterium zbieżności (sumę kwadratów odchyleń od średniej – środka ciężkości)

$$E = \sum_{i=1}^{k} \sum_{x \in C_i} \left| x - m_i \right|^2 \rightarrow \min$$

1) arbitralnie wybierz „k” obiektów jako centra inicjujące skupienia,
2) przydziel (przydziel ponownie) każdy obiekt do skupienia, do którego jest najbardziej podobny, bazując na „średniej wartości obiektów w skupieniu”, czyli średniej odległości od środka ciężkości (*),
3) przelicz (przelicz od nowa) odległość wszystkich obiektów od nowego środka ciężkości dla skupienia (średnią wartość obiektów dla każdego skupienia),

powtarzaj kroki 2)-4) dopóki nie nastąpią żadne zmiany.

## Algorytm PAM (Partitioning Around Medoids)

Algorytm PAM  oparty na wyborze reprezentanta (k-medoids).

Poszukuje „k” skupień w „n” obiektach przez poszukiwanie możliwych reprezentantów medoidów w każdym skupieniu, a wiec obiektów zlokalizowanych najbardziej centralnie w skupieniu. W podejściu tym nadal wykorzystuje się zasadę minimalizacji sumy niepodobieństw między obiektami względem punktów referencyjnych – reprezentantów. W algorytmie PAM zbiór punktów początkowych może być wybrany arbitralnie, można go jednak zautomatyzować. Po wstępnym doborze „k” medoidów algorytm w powtórzeniach próbuje w sposób lepszy dobrać medoidy, zmieniając wybór wstępny z fazy I (analizuje wszystkie możliwe pary obiektów, z których jeden jest medoidą, a drugi nie – w ramach skupienia). Miara jakości grupowania jest liczona dla każdej kombinacji zmiany obiektu. Najlepszy wybór medoid w jednej iteracji jest początkiem kolejnej

iteracji. Dla dużych zbiorów danych jest to algorytm dość kosztowny, pod względem ilości sprawdzanych kombinacji. I faza algorytmu polega na wybraniu „k" reprezentantów, druga dotyczy ich ewentualnych zmian.

Faza I – dobór reprezentantów**:**

- pierwszym reprezentantem $r_1$ jest obiekt o najmniejszej średniej odległości od wszystkich obiektów,

- drugi reprezentant $r_2$ jest wybierany spośród pozostałych obiektów (nie będących reprezentantami), przyjmuje się, że każdy $j-$ ty obiekt może być $r_2$. Dla pierwszego $r_1$ i ustalonego $j$ dokonuje się podziału zbioru obiektów na dwie grupy $k1$ i $k2$ (według mniejszej odległości obiektów od reprezentanta $r_1$ i potencjalnego $j$) i wyznacza się $\bar{d}(j)$, czyli średnią arytmetyczną odległość.

$$\bar{d}(j)=\frac{\sum_{i\in k1}^{n_1} d(i,r_1)+\sum_{i\in k2}^{n_2} d(i,j)}{n}$$

, gdzie, $j$ oraz $r_1$ to reprezentanci, $n_1$ $n_2$ to liczebność grup $k1$ i $k2$.

- reprezentanta wybiera się po sprawdzeniu wszystkich potencjalnych $j-$ tych kandydatów ze zbioru $(n-1)$ obiektów. Drugim reprezentantem będzie taki $j-$ ty obiekt, dla którego wartość $\bar{d}(j)$ jest najmniejsza, zatem:

$$r_2=\min_j \bar{d}(j)$$

Trzeci i następny reprezentant wybierani są podobnie jak drugi. Trzeci, przy ustalonych już wcześniej $(r_1,r_2)$ ze zbioru $(n-2)$ obiektów niebędących reprezentantami.

Faza II – zmiana reprezentantów**:**

W ustalonym zbiorze reprezentantów $\{r_1,...,r_k\}$ i pozostałym zbiorze obiektów $\{j\}$ niebędących reprezentantami zmienia się pary obiektów $\{(r_1, j),...,(r_k, j)\}$, gdzie $j \notin \{r_1,...,r_k\}$. Dla zmienionego zestawu reprezentantów $\{j, r_2,..., r_k\},...,\{r_1, r_2,..., j\}$ dokonuje się nowego podziału zbioru i sprawdza, czy po zmianie $l-$ tego reprezentanta na $j-$ ty obiekt zmniejsza się funkcja celu:

$$F = \frac{\sum_{i=1}^{n} d\left(i^{(l)}, r_l\right)}{n},$$

gdzie $l-$ numer grupy dla $i-$ tego obiektu, $r_l-$ to $l-$ ty reprezentant.

## Algorytm CLARA (Clustering Large Applications)

Typowy algorytm k-medoidów jakim jest algorytm PAM pracuje efektywnie dla niedużego zbioru danych. Do pracy z dużymi zbiorami opracowano jego zmodyfikowaną wersję, algorytm CLARA.

Idea CLARA jest następująca:

Zamiast przeglądać cały zbiór danych weźmy pod uwagę tylko małą jego część (próbę) wybraną w sposób reprezentatywny, a medoidy wybierzmy z niej za pomocą algorytmu PAM. Jeśli próba jest wybrana w sposób losowy i odpowiednio reprezentuje cały zbiór danych, to wybrane w próbie obiekty powinny być podobne do obiektów z całego zbioru danych. CLARA pobiera wielokrotnie próby ze zbioru danych, stosuje PAM dla każdej z prób i daje na wyjściu najlepsze grupowanie dla „k” medoidów – minimalizując średnią odległość „F”.

Efektywność CLARA zależy od rozmiaru próby. Zauważmy, że PAM szuka najlepszych „k” medoidów w danym (całym) zbiorze, a CLARA szuka ich w wylosowanej próbie z tego zbioru. Zatem CLARA nie będzie mogła znaleźć najlepszego grupowania, jeśli

żaden z wybranych z próby medoidów nie będzie wśród faktycznych „k” najlepszych medoidów.
Dobre grupowanie oparte na próbie niekoniecznie będzie reprezentowało dobre grupowanie oparte na całej zbiorowości, jeśli wylosowana próba będzie obciążona.

## Algorytm CLARANS (Clustering Large Applications based on Randomized Search)

W celu poprawy jakości i skalowalności CLARY na cały zbiór obiektów opracowano CLARANS, jest to algorytm typu k-medoidów i kombinacji techniki próbkowania zbioru przy wykorzystaniu algorytmu PAM.
W przeciwieństwie do CLARY, która ma ustaloną próbę losowaną na każdym kroku poszukiwań, CLARANS pobiera próbę z pewną przypadkowością na każdym kroku przeszukiwań. Proces analizy skupień można tutaj porównać do przeszukiwania grafu, gdzie każdy węzeł jest potencjalnym rozwiązaniem tzn. zbiorem k reprezentantów. Grupowanie otrzymane po zamianie 1 medoidy jest nazywane „sąsiadem” obecnego rozwiązania. Liczba sąsiadów, jaka jest losowo „przeszukiwana” jest parametrem Jeśli zostanie znalezione lepsze rozwiązanie, CLARANS przechodzi do tego „sąsiedniego węzła” i proces poszukiwania rozwiązania startuje od nowa, bądź też rozwiązanie to jest lokalnym optimum. Jeśli zostanie znalezione lokalne optimum CLARANS startuje z nowym losowym doborem węzłów w celu poszukania nowego lokalnego optimum.
Eksperymenty pokazały, że jest to algorytm bardziej efektywny niż PAM i CLARA. Złożoność obliczeniowa CLARANS w każdej iteracji jest liniowo proporcjonalne do liczby obiektów. Algorytm ten pozwala wykrywać obserwacje odstające, czyli obiekty nie należące do żadnego ze skupień.