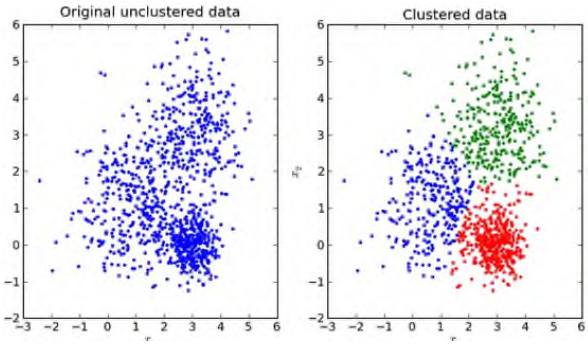
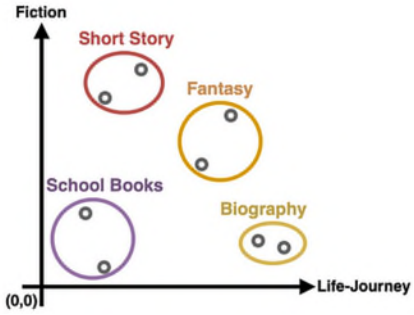


Извор: [2]



Извор: [3]

Кластеризација података: K-means

Кластеризација

Живимо у времену *великих података* (*big data*) које карактерише огромна количина података који се свакодневно прикупљају на најразноврсније начине: био-медицинским истраживањима, мониторингом у пољопривреди, путем сателитских снимака, помоћу камера које прате мобилност и транспорт, аквизицијом и мерењима у производњи базираној на IoT (*Internet of Things*) парадигми и на многе друге начине.

Креирање знања и корисних информација из *big data* универзума превазилази људске способности, те се за стварање знања, поред осталих, користе и методе *кластеризације* (*clustering*). Ове методе и технике припадају домену *рударења података* (*data mining*).

Циљ кластеризације је идентификовање група *сличних* објеката (*similarity objects*), односно препознавање образаца и шаблона (*patterns*) за *интелигентну* поделу или партиционисање скупова података.

Кластеризација припада групи *ненадгледаних* техника *машинског учења* (*unsupervised machine learning*), јер истраживач није вођен неким априорним идејама о томе који објекти (мерења, опсервације…), припадају којој класи.

Кластерска анализа има популарну примену у многим областима: у екологији за партиционисање сличних биолошких регија, у истраживањима рака за груписање пацијената са сличним генским профилима, у маркетингу за сегментацију тржишта, у образовању за профилисање студијских програма оријентисаних ка студентима итд.

Постоји велики број техника за кластеровање [1]:

- партиционо
  - K-Means,
  - K-medoids,
  - CLARA, …
- хијерархијско
  - Agglomerative clustering
  - Divisive clustering…
- напредне технике;
  - базиране на густини (DBSCAN – Density-Based Clustering),
  - Hierarchical K-Means Clustering
  - Fuzzy Clustering
  - Model-Based Clustering итд.

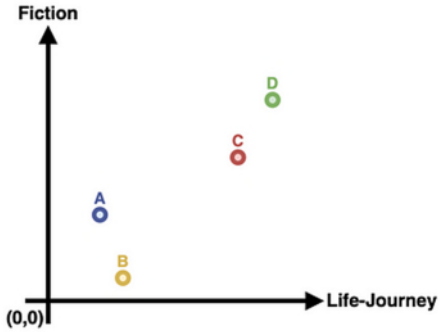
Такође, бројне су и парадигме везане за валидацију кластера као што су: визуелизација, евалуација кластерске тенденције, одређивање оптималног броја кластера, метрике и валидациона статистика кластера…

Мере сличности (Clustering Distance Measures)

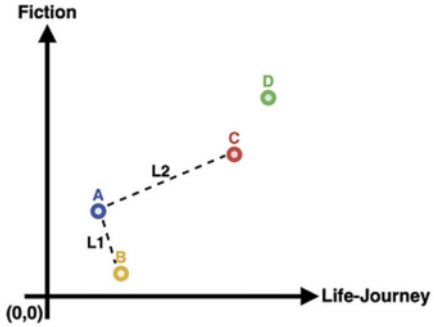
Груписање објеката (опсервација) у кластере захтева израчунавање *дистанце* или *сличности* (*различитости*) (*the dis/similarity distance*) између сваког пара опсервација. Резултат тих прорачуна је *матрица сличности* или *матрица дистанци* (*a dissimilarity or distance matrix*).

Постоји већи број идеја (начина, врста, типова) за одређивање *блискости*. Избор типа блискости (*the choice of distance measures*) је критичан корак у кластеризацији, јер утиче на *облик* генерисаних кластера као и на тумачење добијених резултата [3].

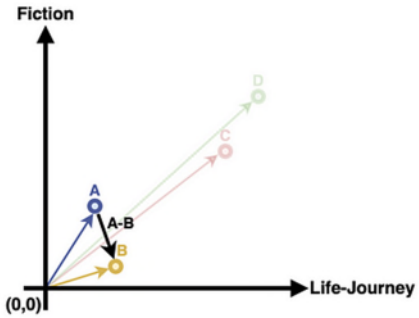
Као уобичајене мере за одређивање блискости користе се: Еуклидова дистанца (*Euclidean distance*) и *косинусна сличност* (*Cosine similarity*) (слике 1- 4) [3].



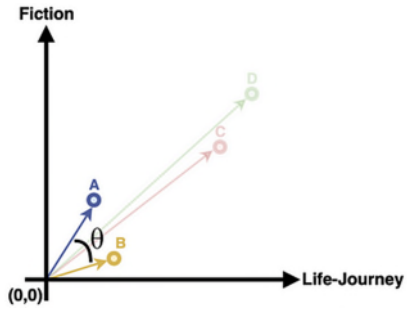
Слика 1. Пример: Четири књиге постављене у 2D коорд. систем [Life-journey, fiction]. Извор: [3]



Слика 2. Књиге A и B имају већу *сличност* него књиге A и C (L1 < L2). Извор: [3]



Слика 3. Еуклидска сличност књиге A и B (A - B). Извор: [3]

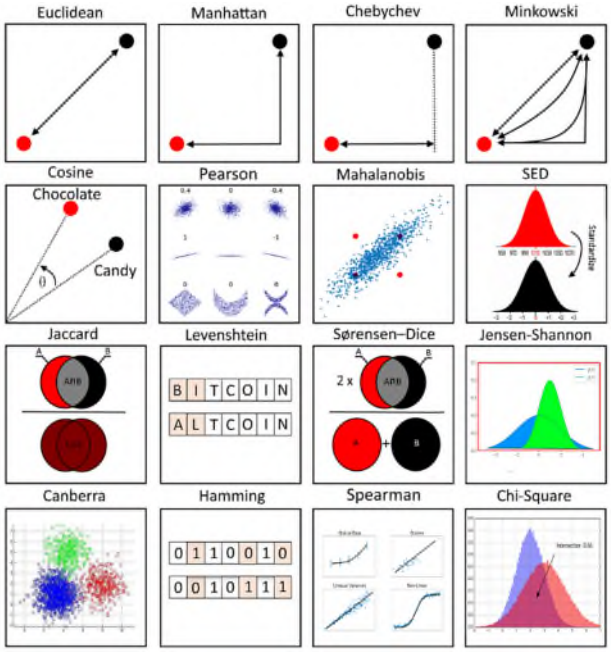


Слика 4. Косинусна сличност књиге A и B (A - B). Извор: [3]

$$Euclidean\ Similarity = \sqrt{\sum_{i=1}^{n}(A_i - B_i)^2}$$

$$Cosine\ Similarity = 1 - \frac{A \cdot B}{\|A\| \cdot \|B\|} = 1 - \frac{\sum_{i=1}^{n} A_i \cdot B_i}{\sqrt{A_i^2} \cdot \sqrt{B_i^2}}$$

Постоје и мере сличности засноване на корелацији као што су: *Pearson correlation distance, Spearman correlation distance, Kendall correlation distance* али и многе друге засноване на различитим концептима: *Manhattan, Gower, Chebychev, Minkowski, Jaccard* и др. (слика 5) [4].



Слика 5. Мере *блискости* у кластер анализи. Извор: [4]

Стандардизација

Мере *сличности* / *блискости* су повезане са скалама на којима се врши мерење.

Да би се избегла доминација појединих варијабли у простору истраживања, најчешће се врши *стандардизација* по свим димензијама. Циљ је да варијабле буду упоредиве.

Генерално, варијабле се скалирају тако да имају: а) стандардну девијацију, $\sigma = 1$ и б) аритметичку средину нула, $\bar{x} = 0$.

При скалирању варијабли (стандардизацији) подаци се трансформишу, на пример, на следећи начин:

$$z_i = \frac{x_i - center(x)}{scale(\sigma)} = \frac{x_i - \bar{x}}{\sigma}.$$

K-means clustering

K-means кластеризација (*MacQueen, 1967*) је најчешће коришћена метода кластеризације. У овој методи *жељени* број кластера (k), унапред задаје истраживач. Метода класификује објекте (мерења, опсервације,..) у више група (кластера) тако да објекти унутар кластер буду што блискији (*as similar as possible*), односно тежи се *што већој унутаркластерској сличности* (*high intra-class similarity*), док различитост међу објектима различитих класа треба да буде што већа (*as dissimilar as possible*), односно тежи се *што мањој међукластерској сличности* (*low inter-class similarity*).

У *k-means clustering*-у, сваки кластер је репрезентован својим *центроидом* (*centroid*) који је дефинисан средњом вредношћу тачака које су придружене кластеру [1].

K-means базична идеја

Главна идеја *k-means* алгоритма кластеризације је минимизација **укупне унутаркластерске варијације** (*total within-cluster variation*). Постоји више алгоритама за ову минимизацију али је највише у употреби стандардни *Hartigan-Wong* алгоритам (1979), који дефинише укупну унутаркластерску варијацију као суму квадрата Еуклидских дистанци сваког члана кластера до одговорајућег центроида:

$$tot.withiness = \sum_{i=1}^{k} W(C_k) = \sum_{i=1}^{k} \sum_{x_i \in C_k} (x_{k,i} - \mu_k)^2$$

У претходном изразу $x_{k,i}$ је тачка која припада кластеру C_k, док је μ_k је centroid кластера C_k.

Метрика *total within-cluster sum of square* мери *компактност* (*goodness*) кластеризације, а циљ је да ова метрика буде што је могуће мања.

K-means алгоритам

K-means алгоритам се може сажети у следеће кораке [1]:

1. Одредити број кластера (k) који ће бити креиран (од стране аналитичара);
2. Одабрати насумично центре будућих кластера;
3. Доделити сваку опсервацију (мерење, објекат…) кластеру чији центар је *најближи*, према одабраној *мери сличности*;
4. За сваки од k кластера, ажурирати цетре кластера (*центроиде*), израчунавањем нове позиције центроида, на основу свих тачака (опсервација) у посматраном кластеру;
5. Итеративно минимизовати *укупну унутаркластерску варијацију* (*tot.withiness*), док промене позиције центроида не постану мање од неког задатог ε, или док се не дође до неког задатог максималног броја итерација.

Пример k-means кластеризације у R-u

Дат је пример са коментарима *k-means* кластеризације, за анонимизовани скуп података који се односи на студенте студијског програма *Информационе технологије* **Академије Западна Србија**, Одсек у Ужицу и њихове исходе учења током периода: 2011-2019. год. Скуп садржи податке за 424 студента, и следеће варијабле: ***y1_mg*** *– просечна оцена након прве године студија,* ***y1_espb*** *– број ЕСПБ након прве године студија,* ***usp_ss*** *– успех из средње школе,* ***usp_mbe*** *– успех на пријемном испиту из математике,* ***usp_tk*** *– успех на пријемном испиту на тесту културе, као и оцене:* ***Baze*** *– Базе података,* ***Mat1*** *– Математика 1,* ***ElElek*** *– Електротехника са електроником, OS – Оперативни системи,* ***Mat2*** *– Математика 2,* ***Engl1*** *– Енглески 1,* ***RU*** *– Рачунарско управљање,* ***AlgSP*** *– Алгоритми и структуре података.*

```
> # Instalacija i ucitavanje potrebnih paketa
> install.packages("factoextra")
> library(factoextra)
> install.packages("readxl")
> library(readxl)
```

```
> # Ucitavanje podataka i prikaz prva 3 zapisa
> informatika <- read_excel("d://Blog23mm//IN_11-19_Blog.xlsx")
> head(informatika, n=3)
# A tibble: 3 x 13
  y1_mg y1_espb usp_ss usp_mbe usp_tk  Baze  Mat1 ElElek    OS  Mat2 Engl1    RU AlgSP
  <dbl>   <dbl>  <dbl>   <dbl>  <dbl> <dbl> <dbl>  <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
1  7.78      54   22.0    18.5   30       8     7      8     9     7     5     8     5
2  8.5       60   30.1    30     28.5    10     7     10    10     8     7     9     5
3  7.57      42   22.5     9.5   22.5     5     5      8     7     5     5     5     5
```

```
> # Izbor podskupa od 50 slucajno odabranih studenata i prikaz prva 3 zapisa iz podskupa
```

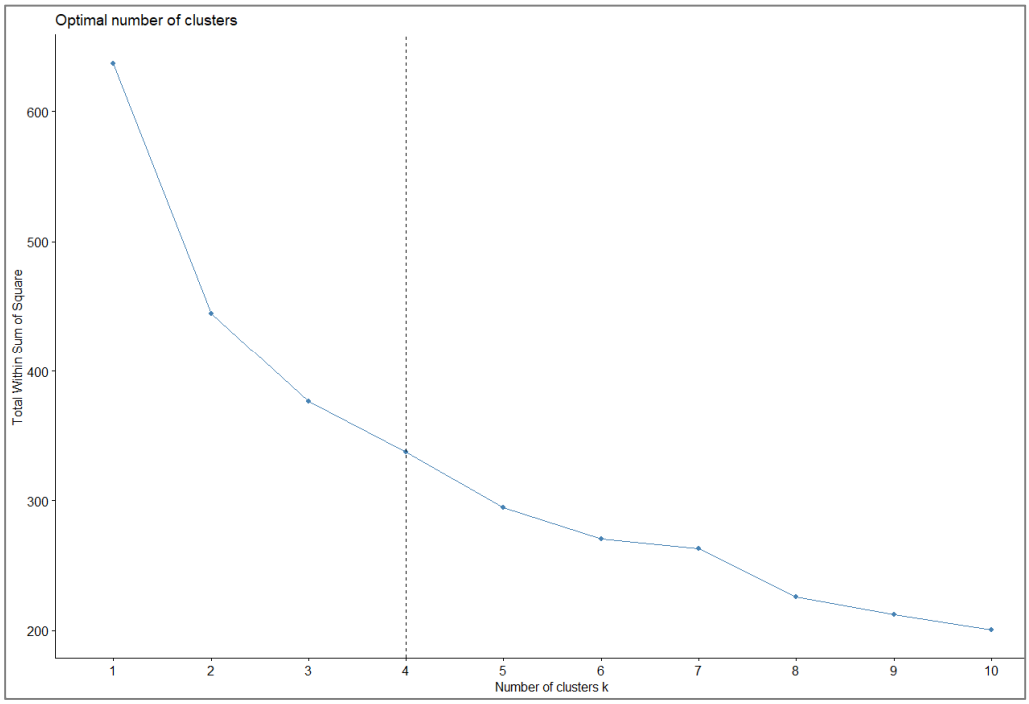
```
> # Podskup podataka
> set.seed(123)
> ss <- sample(1:442, 50) # Preuzimanje 50 slicajnih zapisa od ukupno 442
> df <- informatika[ss, ] # podskup od 50 zapisa
> head(df,n=3)
# A tibble: 3 x 13
  y1_mg y1_espb usp_ss usp_mbe usp_tk  Baze  Mat1 ElElek    OS  Mat2 Engl1    RU AlgSP
  <dbl>   <dbl>  <dbl>   <dbl>  <dbl> <dbl> <dbl>  <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
1  7.89      54   32.4      18   19.5     8     6      7     7     6    10     5     9
2  8.1       60   35.6      12   27       8     7      8     9     6     5     5     7
3  7.4       60   29.7      23.5 22.5     7     7      8     8     6     5     8     5
```

```
> df.scaled <- scale(df) # Standardizacija varijabli

> #Izracunavanje Euklidske distance
> dist.eucl <- dist(df.scaled, method = "euclidean")

> # Reformatiranje podataka u formu matrice, podskup prvih 6 kolona i redova i zaokrizivanje vrednosti
> round(as.matrix(dist.eucl)[1:6, 1:6], 2) #  nije prikazano zbog obima

> # Odredivanje optimalnog broja klastera pomocu Scree Plot-a
> fviz_nbclust(df.scaled, kmeans, method = "wss") +
+     geom_vline(xintercept = 4, linetype = 2)
```



Слика 6. Оцена оптималног броја кластера: к = 4 кластера. Извор: аутор

```
> # realizacija klasterovanja: 4 klastera
> set.seed(123)
> km.res <- kmeans(df.scaled, 4, nstart = 25)
> # prikaz rezultata klasterovanja
> print(km.res) # nije dato zbog obima

> # prikaz pozicije centroida (reprezenata klastera) u prirodnim koordinatama:  4 klastera
> aggregate(df, by=list(cluster=km.res$cluster), mean) #u originalnim koordinatama
  cluster    y1_mg  y1_espb   usp_ss  usp_mbe   usp_tk     Baze     Mat1   ElElek       OS     Mat2    Engl1       RU     AlgSP
1       1 6.684430 46.94118 24.31882 15.76471 23.55882 5.529412 6.058824 6.411765 6.588235 5.823529 6.000000 5.529412 5.529412
2       2 9.000000 60.00000 30.04667 25.00000 29.50000 10.000000 8.666667 8.333333 10.000000 7.666667 9.666667 10.000000 10.000000
3       3 6.964286 15.42857 28.46857 19.85714 25.07143 5.000000 5.000000 5.714286 5.285714 5.000000 5.428571 5.000000 5.000000
4       4 7.445498 54.26087 32.28261 21.89130 22.30435 7.869565 6.565217 7.695652 8.000000 6.434783 6.826087 6.521739 6.956522
> # u prikazu iznad moze se uociti da klaster br. 2 predstavlja najbolje studente
```

> # u nastavku je data klasterska pripadnost za svakog sudenta, kao i broj studenata po klasterima

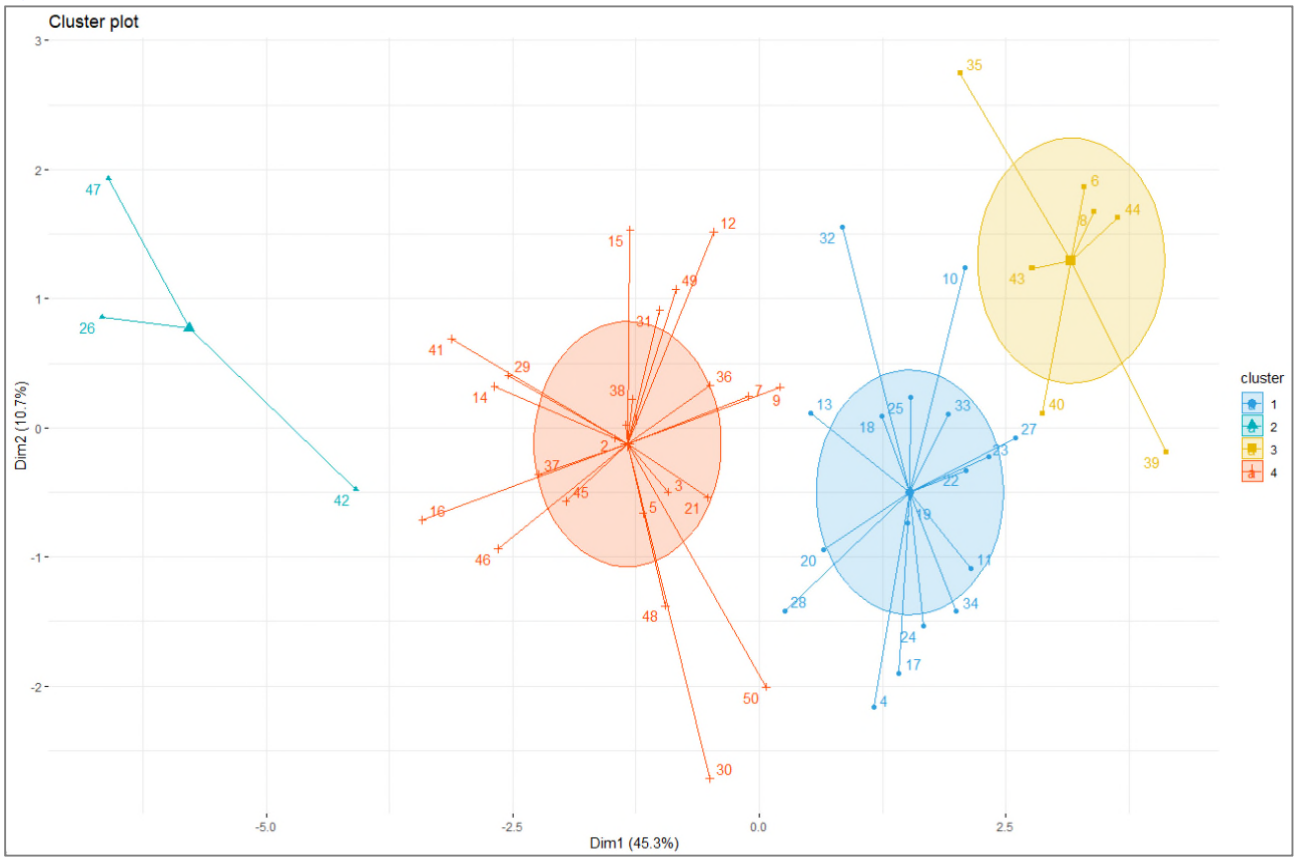
```
> # Cluster number for each of the observations
> km.res$cluster
 [1] 4 4 4 1 4 3 4 3 4 1 1 4 1 4 4 4 1 1 1 1 4 1 1 1 1 2 1 1 4 4 4 1 1 1 3 4 4 4 3 3 4 2 3 3 4 4 2 4 4 4
> # Cluster size
> km.res$size
[1] 17  3  7 23
```

Резултати поступка кластеризације могу се визуелизовати, што може бити корисно за процену правилности у избору броја кластера. За визуелизацију је погодан дијаграм распршености (*scatter plot*) са бојењем тачака, бојама које означавају кластерску припадност.

Пошто је проблем 13-то димензионалан (сваки студент је представљен са 13 варијабли), погодно је смањити број димензија погодним алгоритмом, као што је алгоритам *Анализа главних компоненти* (***PCA** – Principal Component Analysis*). PCA алгоритам, 13 оригиналних варијабли *редукује* на **две главне компоненте** које се могу приказати у новом 2D простору.

Функција **fviz_cluster()** [faktoekstra пакет] се може користити за лаку визуелизацију k-means кластера. У резултујућем дијаграму, опсервације су представљене тачкама са координатама у форми главних компоненти [*Dim1*, *Dim2*]. Такође је могуће за сваки сваки кластер креирати *елипсу концентрације* [1].

```
> fviz_cluster(km.res, data = df.scaled,
+              palette = c("#2E9FDF", "#00AFBB", "#E7B800", "#FC4E07"),
+              ellipse.type = "euclid", # elipse koncentracije
+              star.plot = TRUE, # dodatvanje linijskih segmenata od centroida do opservacije/ studenta
+              repel = TRUE, # izbegavanje precrtavanja oznaka
+              ggtheme = theme_minimal()
+ )
```



Слика 7. Приказ кластера студената након редукције димензионалности на две главне компоненте помоћу PCA алгоритма /*студенти су представљени бројевима због заштите и приватности података*/. Извор: аутор

Литература

[1] Kassambara A. (2017), Multivariate Analysis I, Practical Guide To Cluster Analysis in R, Unsupervised Machine Learning, Published by STHDA.

[2] https://towardsdatascience.com/understanding-the-concept-of-hierarchical-clustering-technique-c6e8243758ec

[3] https://towardsdatascience.com/introduction-to-embedding-clustering-and-similarity-11dd80b00061

[4] https://towardsdatascience.com/17-types-of-similarity-and-dissimilarity-measures-used-in-data-science-3eb914d2681

[5] Milivojevic M., Forst Dj., Stopic S., Drndarevic D., Stevanovic M., (2014), Development of software for k-means clustering, 7. Int. Conf. SED 2014, Uzice, Serbia, Oct. 2014. (2.10-2.17), ISBN 978-86-83573-42-24, COBIS.SR-ID 209983756

[6] Milivojevic M., Forst Dj., Subotic Lj., Djokovic K., (2016), Assesment of row water quality by using k-means clustering software, 9. Int. Conf. SED 2016, Uzice, Serbia, 30. Sep. – 01. Oct. 2016. (2.53-2.63), ISBN 978-86-83573-82-02, COBIS.SR-ID 227527948

Аутор,
др Милован Миливоејвић, проф. стр. ст.