Úloha 5 – K-Means algoritmus

Zadání:

1. Doimplementujte K-Means algoritmus. Přiložená funkce v Matlabu implementuje část K-Means algoritmu (nalezení nejbližších reprezentantů (centroidů)). Vaším úkolem je doplnit přesun reprezentantů do středu nových shluků a určit, zda je možné ukončit algoritmus nebo má smysl pokračovat další iterací.
2. Centroidy (reprezentanty) inicializujte náhodně a při každém spuštění jinak.
3. Shlukněte přiložená data vaším K-Means algoritmem. Zkuste různé počty reprezentantů (2, 3, … 10). Spočítejte průměrnou siluetu pro všechny počty shluků a určete, pro který počet reprezentantů vyjde průměrná silueta nejlépe. Pro zajímavé počty reprezentantů zobrazte grafy siluet.
4. Pro nejlepší počet reprezentantů, který vám vyšel v minulém bodě, (alespoň 5x) spusťte algoritmus K-Means s různými náhodnými počátečními pozicemi reprezentantů.
5. Shlukněte data pomocí hierarchického shlukování. Vytvořte stejný počet shluků, který vám vyšel nejlépe v algoritmu K-Means. Do zprávy vložte dendrogram, graf siluety a průměrnou siluetu. Krátce okomentujte rozdíly mezi výsledky hierarchického shlukování a K-Means algoritmu.

Užitečné funkce: silhouette, kmeans, linkage, pdist, cluster, cophenet, scatter

Obsah zprávy:

1. Doplněný zdrojový kód. A jeho stručný popis.

```
function [nearestIdx, coordinates] = myKMeans(vectors, centroids)
% KMeans algoritmus.
%
% vectors - instance/data, ktera se maji shlukovat. (Matice M x N - radky
% jednotlive vektory, sloupce dimenze)
% centroids - pocatecni nastaveni centroidu (Matice K x N)
%
% nearestIdx - vysledne prirazeni vektoru k jednotlivym centroidum
% coordinates - vysledne pozice centroidu (reprezentantu)

numVec = size(vectors,1);

stopping = false;
while(~stopping)

    %Vypocet nejblizsiho reprezentanta
    nearestIdx = zeros(numVec, 1);
    for i = 1:numVec
        nearestIdx(i) = nearestCentroid(vectors(i,:), centroids);
    end

    %Presun centroidu (reprezentantu)
    %Tuto cast mate doplnit

clusterCount = size(centroids, 1);
      %pocet clusteru = veliksot pole, ktere jde do algoritmu
    arrClusterSize = zeros(clusterCount, 1);
      % pole, do ktereho se uklada, kolik maji shluky prvku
      % na zacatku je tam 0 prvku

    for i = 1:numVec
      arrClusterSize(nearestIdx(i),1) = arrClusterSize(nearestIdx(i),1)+1;
      % projde vsechny clustery
    end

    tempCentroids = zeros(clusterCount, size(vectors,2));
```

Zpracovala: **Radoslava Jandová** E-mail: **jandora1@fel.cvut.cz**
Obor: **STM, Softwarové inženýrství** Semestr: **zimní 2011/2012**

```
    for i = 1:clusterCount
      tempCentroids(i,:) = 1/arrClusterSize(i,1)
            *sum(vectors(nearestIdx==i,:))
        % projde vsechny centroidy a vytvori nove lokace
        % tj. secte vsechny vektory daneho shluku
        % nearestIndx==i - vytvori nulovy vektor, ktery neovlivni vysledek
        % je to pro pripad, ze vychazi index v poli 0
    end

    %Mam pokracovat?
    %Tuto cast mate doplnit

    iterationCount = iterationCount + 1;
        % zvysi se pocet iteraci
    if (isequalwithequalnans(centroids,tempCentroids))
        % isequalwithequalnans = fce matlabu, ktera porovna matice
        % a predpoklada, ze NaN (= not a number) jsou stejne "hodnoty"
        % NaN vznikne, pokud v clusteru neni zadny prvek

    stopping = true;
end

coordinates = centroids;

end

function nearestCentrIdx = nearestCentroid(instance, centroids)
% Vypocet eukleidovske vzdalenosti mezi jednim vektorem a vsemi centroidy
% (reprezentanty).
% instance - vektor
% centroids - centroidy
% Vraci index nejblizsiho centroidu.

    numCentroids = size(centroids, 1);
    x = repmat(instance, numCentroids, 1);

    dist = x - centroids;
    dist = dist .* dist;

    dist = sqrt(sum(dist,2));

    [tmp, nearestCentrIdx] = min(dist);

end
```

2. Průměrné hodnoty siluety pro počty reprezentantů: 2, 3, 4, ... 10.

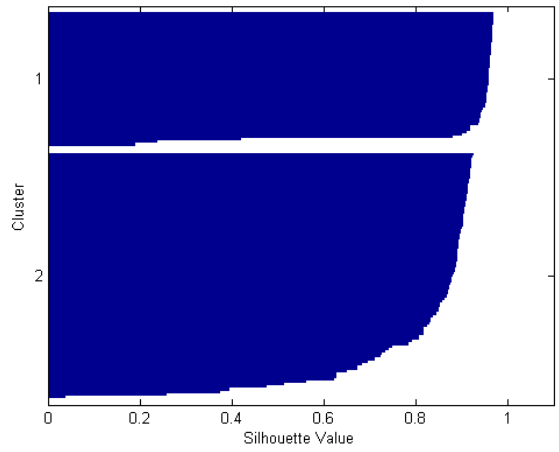
V následujícím přehledu uvádím přehledy při rozdělení na 2, 3, 4, ... 10 shluků. Přehledy obsahují
- počet shluků
- počet iterací
- hodnoty počátečních souřadnic
- hodnoty výsledných souřadnic
- 2D graf

Pro 2 reprezentanty:

Počet iterací: 7

Počáteční centroidy:
C1 = 6.7000 3.0000 5.2000 2.3000
C2 = 6.8000 3.2000 5.9000 2.3000

Konečné centroidy:
C1 = 5.0057 3.3698 1.5604 0.2906
C2 = 6.3010 2.8866 4.9588 1.6959

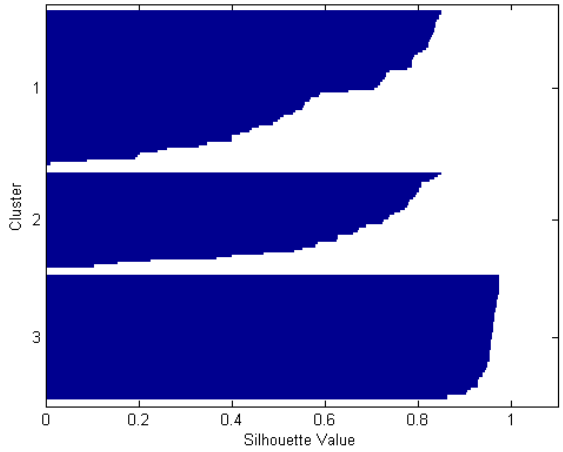


Pro 3 reprezentanty:

Počet iterací: 4

Počáteční centroidy:
C1 = 6.3000 2.5000 4.9000 1.5000
C2 = 6.9000 3.2000 5.7000 2.3000
C3 = 5.1000 3.7000 1.5000 0.4000

Konečné centroidy:
C1 = 5.9016 2.7484 4.3935 1.4339
C2 = 6.8500 3.0737 5.7421 2.0711
C3 = 5.0060 3.4280 1.4620 0.2460

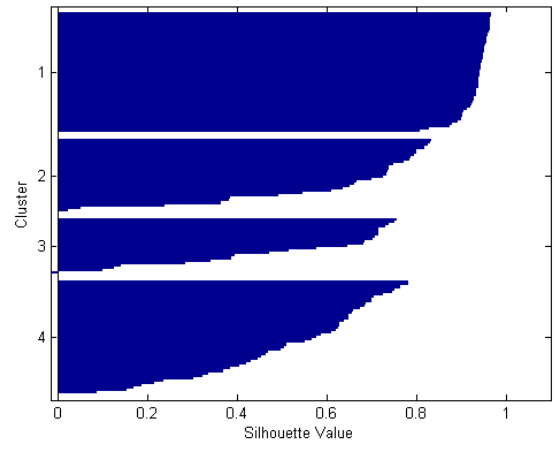


Pro 4 reprezentanty:

Počet iterací: 18

Počáteční centroidy:
C1 = 6.1000 2.9000 4.7000 1.4000
C2 = 6.4000 3.1000 5.5000 1.8000
C3 = 7.7000 2.6000 6.9000 2.3000
C4 = 6.8000 3.2000 5.9000 2.3000

Konečné centroidy:
C1 = 5.0060 3.4280 1.4620 0.2460
C2 = 5.5800 2.6333 3.9867 1.2333
C3 = 7.0870 3.1261 6.0130 2.1435
C4 = 6.2936 2.9000 4.9511 1.7298



Zpracovala**: Radoslava Jandová**
Obor: **STM, Softwarové inženýrství**
E-mail: **jandora1@fel.cvut.cz**
Semestr: **zimní 2011/2012**

Pro 5 reprezentantů:

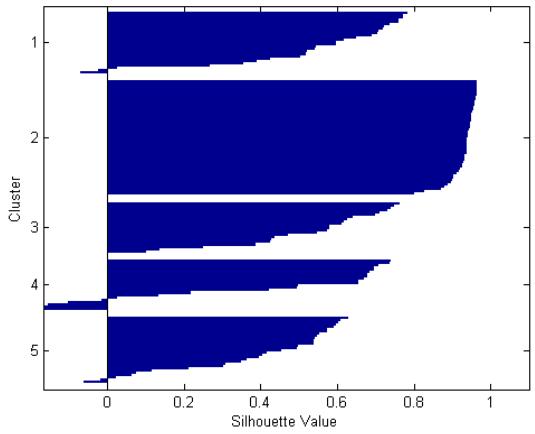
Počet iterací: 8

Počáteční centroidy:

C1 = 5.1000	2.5000	3.0000	1.1000
C2 = 5.4000	3.9000	1.7000	0.4000
C3 = 6.1000	3.0000	4.9000	1.8000
C4 = 6.7000	3.1000	5.6000	2.4000
C5 = 5.8000	2.7000	5.1000	1.9000

Konečné centroidy:

C1 = 5.5185	2.6222	3.9519	1.2185
C2 = 5.0060	3.4280	1.4620	0.2460
C3 = 6.4000	2.9227	4.5864	1.4409
C4 = 7.1227	3.1136	6.0318	2.1318
C5 = 6.1966	2.8828	5.1828	1.9345



Pro 6 reprezentantů:

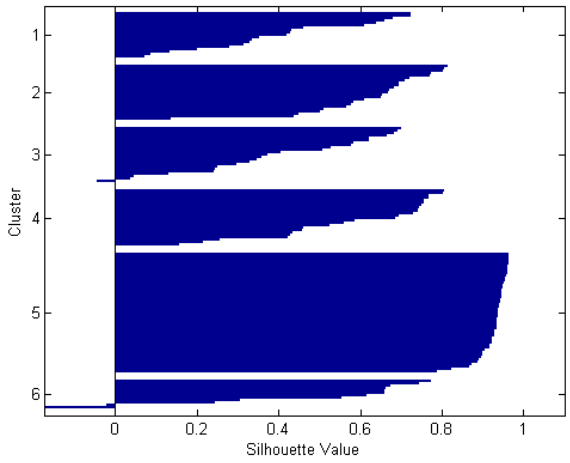
Počet iterací: 7

Počáteční centroidy:

C1 = 5.7000	2.5000	5.0000	2.0000
C2 = 6.4000	2.7000	5.3000	1.9000
C3 = 6.6000	2.9000	4.6000	1.3000
C4 = 5.1000	2.5000	3.0000	1.1000
C5 = 5.0000	3.0000	1.6000	0.2000
C6 = 7.6000	3.0000	6.6000	2.1000

Konečné centroidy:

C1 = 5.9526	2.7632	4.9579	1.7579
C2 = 6.5609	3.0696	5.5261	2.1522
C2 = 6.3609	2.9304	4.5435	1.4261
C3 = 5.4870	2.5739	3.8783	1.1870
C4 = 5.0060	3.4280	1.4620	0.2460
C5 = 7.4750	3.1250	6.3000	2.0500



Pro 7 reprezentantů:

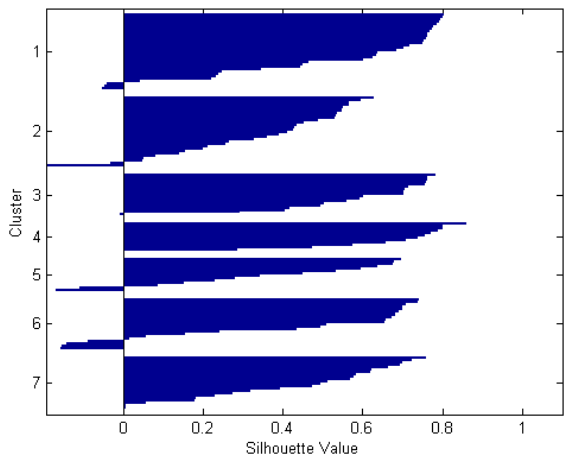
Počet iterací: 10

Počáteční centroidy:

C1 = 4.6000	3.2000	1.4000	0.2000
C2 = 5.8000	2.7000	5.1000	1.9000
C3 = 5.4000	3.9000	1.7000	0.4000
C4 = 6.7000	3.1000	4.4000	1.4000
C5 = 4.9000	2.4000	3.3000	1.0000
C6 = 5.8000	2.8000	5.1000	2.4000
C7 = 6.0000	2.2000	5.0000	1.5000

Konečné centroidy:

C1 = 4.8094	3.2281	1.4344	0.2281
C2 = 6.2000	2.8700	5.1733	1.9200
C3 = 5.3556	3.7833	1.5111	0.2778
C4 = 6.6333	3.0333	4.6333	1.4583
C5 = 5.3571	2.4429	3.7143	1.1643
C6 = 7.1227	3.1136	6.0318	2.1318
C7 = 5.8591	2.8182	4.3227	1.3318



Zpracovala**: Radoslava Jandová**
Obor: **STM, Softwarové inženýrství**
E-mail: **jandora1@fel.cvut.cz**
Semestr: **zimní 2011/2012**

Pro 8 reprezentantů:

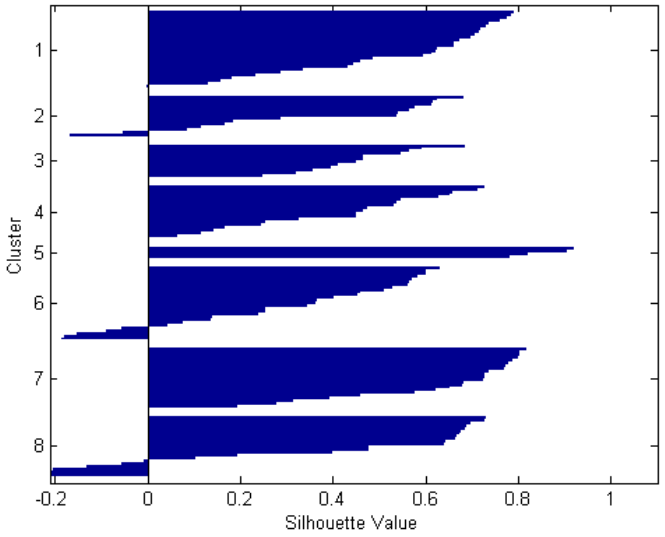
Počet iterací: 8

Počáteční centroidy:

C1 = 5.2000	3.4000	1.4000	0.2000
C2 = 6.1000	2.8000	4.7000	1.2000
C3 = 5.6000	3.0000	4.5000	1.5000
C4 = 5.7000	2.9000	4.2000	1.3000
C5 = 4.9000	2.5000	4.5000	1.7000
C6 = 5.7000	2.5000	5.0000	2.0000
C7 = 4.5000	2.3000	1.3000	0.3000
C8 = 5.8000	2.7000	5.1000	1.9000

Konečné centroidy:

C1 = 5.2429	3.6679	1.5000	0.2821
C2 = 6.5600	2.8933	4.6267	1.4533
C3 = 5.9167	2.9917	4.5500	1.4833
C4 = 5.6150	2.6400	4.0050	1.2250
C5 = 5.0000	2.3000	3.2750	1.0250
C6 = 6.2148	2.8667	5.2111	1.9444
C7 = 4.7045	3.1227	1.4136	0.2000
C8 = 7.1227	3.1136	6.0318	2.1318



Pro 9 reprezentantů:

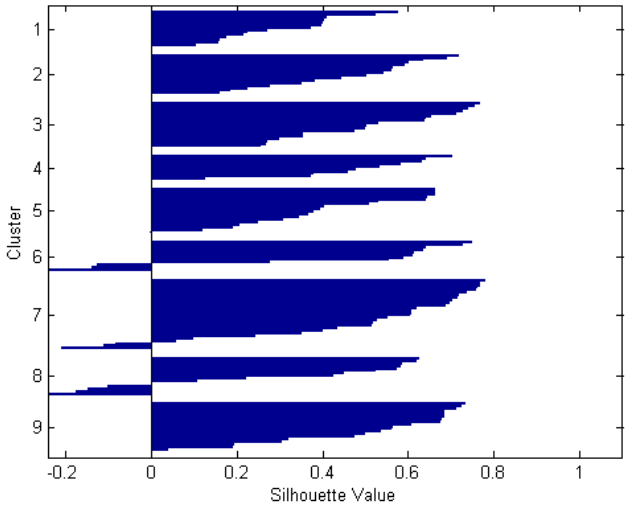
Počet iterací: 6

Počáteční centroidy:

C1 = 6.8000	3.0000	5.5000	2.1000
C2 = 4.4000	3.0000	1.3000	0.2000
C3 = 6.6000	3.0000	4.4000	1.4000
C4 = 6.5000	3.0000	5.8000	2.2000
C5 = 6.3000	2.8000	5.1000	1.5000
C6 = 6.8000	3.2000	5.9000	2.3000
C7 = 5.8000	2.7000	3.9000	1.2000
C8 = 5.4000	3.4000	1.7000	0.2000
C9 = 4.6000	3.6000	1.0000	0.2000

Konečné centroidy:

C1 = 6.5786	2.9429	5.3857	1.9929
C2 = 4.6687	3.0250	1.4125	0.1938
C3 = 6.4278	2.9778	4.5722	1.4167
C4 = 6.5500	3.2400	5.6700	2.3300
C5 = 6.0167	2.7056	4.9833	1.7722
C6 = 7.4750	3.1250	6.3000	2.0500
C7 = 5.5321	2.6357	3.9607	1.2286
C8 = 5.4000	3.8267	1.5200	0.2733
C9 = 4.9789	3.4526	1.4579	0.2684



Zpracovala**: Radoslava Jandová**
Obor: **STM, Softwarové inženýrství**
E-mail: **jandora1@fel.cvut.cz**
Semestr: **zimní 2011/2012**

Pro 10 reprezentantů:

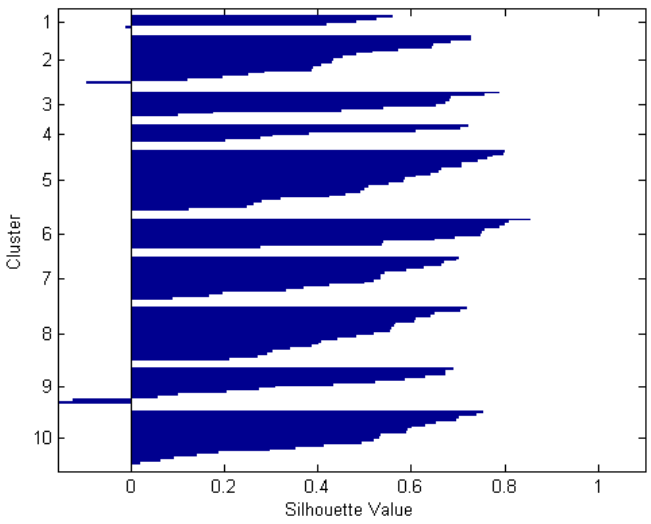
Počet iterací: 13

Počáteční centroidy:

C1 = 4.4000	3.0000	1.3000	0.2000
C2 = 6.2000	2.8000	4.8000	1.8000
C3 = 7.7000	2.8000	6.7000	2.0000
C4 = 5.5000	3.5000	1.3000	0.2000
C5 = 6.9000	3.1000	5.4000	2.1000
C6 = 6.9000	3.1000	4.9000	1.5000
C7 = 4.7000	3.2000	1.6000	0.2000
C8 = 4.9000	3.6000	1.4000	0.1000
C9 = 5.8000	2.6000	4.0000	1.2000
C10 = 5.9000	3.2000	4.8000	1.8000

Konečné centroidy:

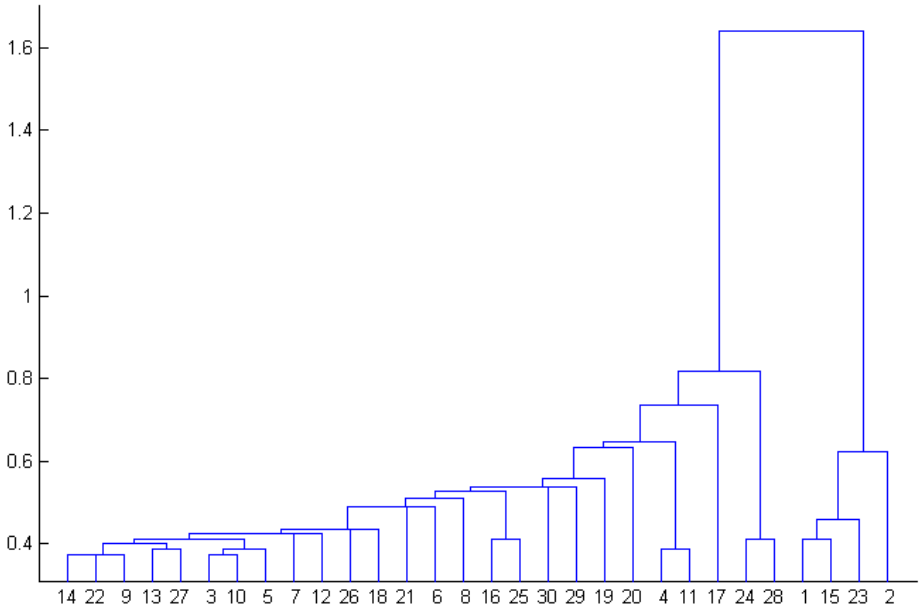
C1 = 4.4000	2.8800	1.2800	0.2000
C2 = 6.0368	2.7053	5.0000	1.7789
C3 = 7.5400	3.1500	6.3900	2.0900
C4 = 5.5286	4.0429	1.4714	0.2857
C5 = 6.6167	3.0792	5.5625	2.1375
C6 = 6.6333	3.0333	4.6333	1.4583
C7 = 4.7941	3.1941	1.4529	0.2000
C8 = 5.1476	3.5429	1.5095	0.2810
C9 = 5.3571	2.4429	3.7143	1.1643
C10 = 5.8429	2.8476	4.3143	1.3238



Z provedených testů je patrné, že nejlepší výsledky vyšly pro 2 a 3 reprezentanty, tzn., že silueta je celkem dobré struktury. U ostatních testů silueta zasahuje již do záporné části grafu, což znamená, že taková silueta rozdělení je méně pravděpodobná.

Dendrogram

```
function dendrogramPresent( dataFileName )
data = csvread(dataFileName);
linkRes = linkage(data);
dendrogram(linkRes)
end
```



Zpracovala: **Radoslava Jandová** E-mail: **jandora1@fel.cvut.cz**
Obor: **STM, Softwarové inženýrství** Semestr: **zimní 2011/2012**

Test siluety pro 2 a 3 shluky

Siluety pro 2 a 3 shluky vyšly v prvním testování jako nejlepší, a proto jsem provedla opakovaná měření pro tyto shluky.

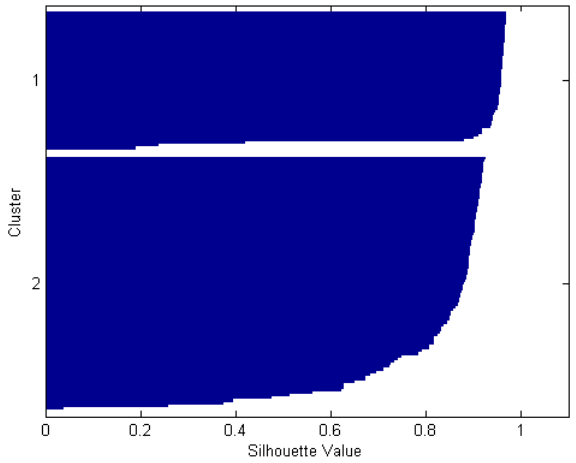
Pro 2 reprezentanty:

1. měření

Počet iterací: 7

Počáteční centroidy:
C1 = 6.7000 3.0000 5.2000 2.3000
C2 = 6.8000 3.2000 5.9000 2.3000

Konečné centroidy:
C1 = 5.0057 3.3698 1.5604 0.2906
C2 = 6.3010 2.8866 4.9588 1.6959

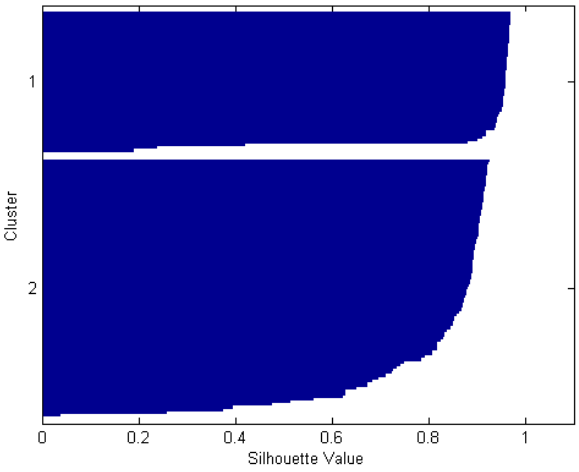


2. měření

Počet iterací: 6

Počáteční centroidy:
C1 = 6.9000 3.1000 4.9000 1.5000
C2 = 6.7000 3.3000 5.7000 2.1000

Konečné centroidy:
C1 = 5.0057 3.3698 1.5604 0.2906
C2 = 6.3010 2.8866 4.9588 1.6959

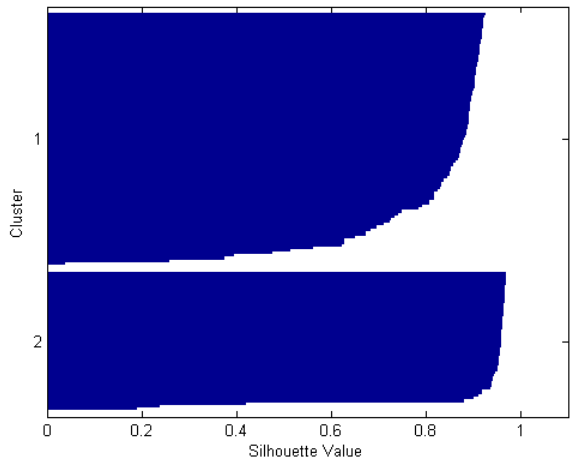


3. měření

Počet iterací: 5

Počáteční centroidy:
C1 = 6.3000 2.3000 4.4000 1.3000
C2 = 5.8000 2.7000 3.9000 1.2000

Konečné centroidy:
C1 = 6.3010 2.8866 4.9588 1.6959
C2 = 5.0057 3.3698 1.5604 0.2906



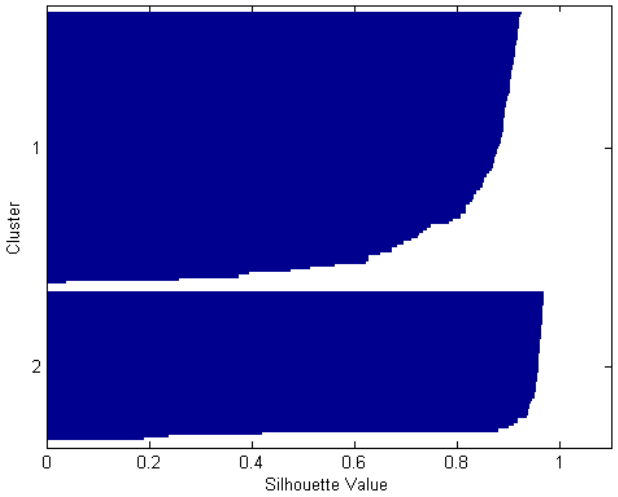
Zpracovala**: Radoslava Jandová**
Obor: **STM, Softwarové inženýrství**
E-mail: **jandora1@fel.cvut.cz**
Semestr: **zimní 2011/2012**

4. měření

Počet iterací: 3

Počáteční centroidy:
C1 = 6.4000 3.1000 5.5000 1.8000
C2 = 4.4000 3.2000 1.3000 0.2000

Konečné centroidy:
C1 = 6.3010 2.8866 4.9588 1.6959
C2 = 5.0057 3.3698 1.5604 0.2906

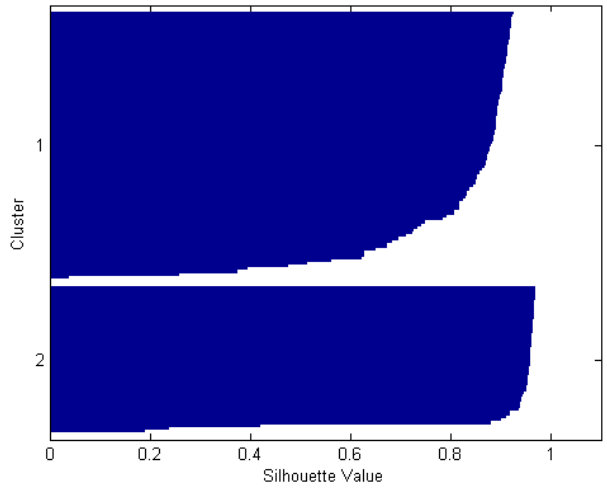


5. měření

Počet iterací: 4

Počáteční centroidy:
C1 = 5.7000 2.5000 5.0000 2.0000
C2 = 4.9000 2.4000 3.3000 1.0000

Konečné centroidy:
C1 = 6.3010 2.8866 4.9588 1.6959
C2 = 5.0057 3.3698 1.5604 0.2906



Z opakovaných testů je patrné, že silueta pro 2 shluky je stále dobré struktury. Počáteční hodnoty centroidů jsou u každého měření jiné, což odpovídá náhodnému výběru. Konečné hodnoty, ale vychází u všech měření stejně, což potvrzuje stabilní strukturu.

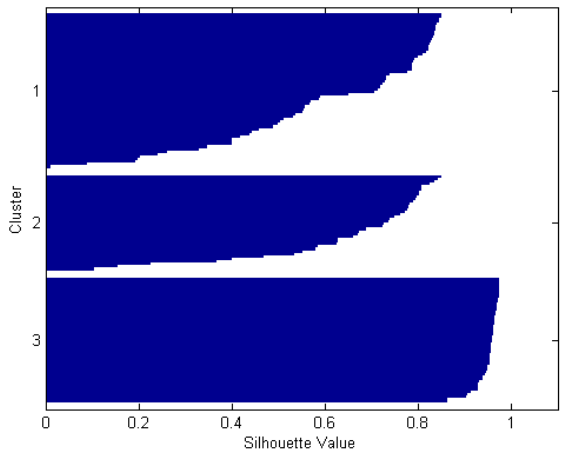
Pro 3 reprezentanty:

1. měření

Počet iterací: 4

Počáteční centroidy:
C1 = 6.3000 2.5000 4.9000 1.5000
C2 = 6.9000 3.2000 5.7000 2.3000
C3 = 5.1000 3.7000 1.5000 0.4000

Konečné centroidy:
C1 = 5.9016 2.7484 4.3935 1.4339
C2 = 6.8500 3.0737 5.7421 2.0711
C3 = 5.0060 3.4280 1.4620 0.2460



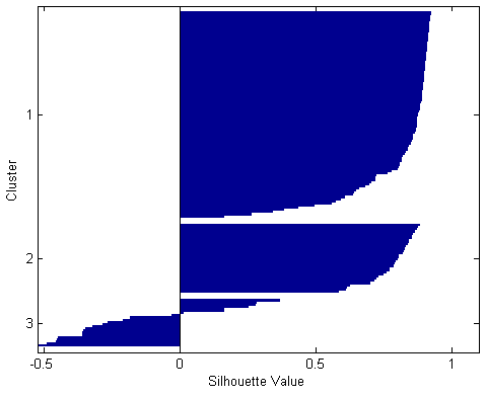
Zpracovala**: Radoslava Jandová** E-mail: **jandora1@fel.cvut.cz**
Obor: **STM, Softwarové inženýrství** Semestr: **zimní 2011/2012**

2. měření

Počet iterací: 7

Počáteční centroidy:
C1 = 6.0000 3.4000 4.5000 1.6000
C2 = 4.8000 3.4000 1.6000 0.2000
C3 = 4.4000 2.9000 1.4000 0.2000

Konečné centroidy:
C1 = 6.3146 2.8958 4.9740 1.7031
C2 = 5.1937 3.6312 1.4750 0.2719
C3 = 4.7318 2.9273 1.7727 0.3500

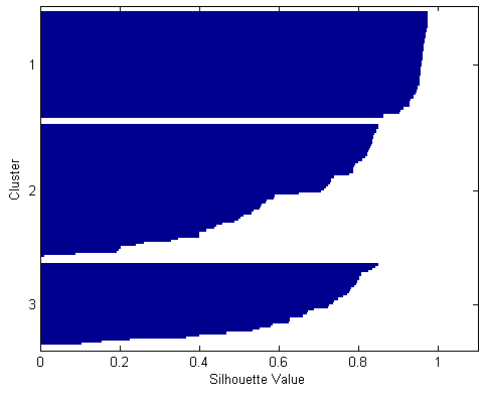


3. měření

Počet iterací: 6

Počáteční centroidy:
C1 = 5.7000 2.6000 3.5000 1.0000
C2 = 6.5000 3.0000 5.5000 1.8000
C3 = 6.7000 3.1000 5.6000 2.4000

Konečné centroidy:
C1 = 5.0060 3.4280 1.4620 0.2460
C2 = 5.9016 2.7484 4.3935 1.4339
C3 = 6.8500 3.0737 5.7421 2.0711

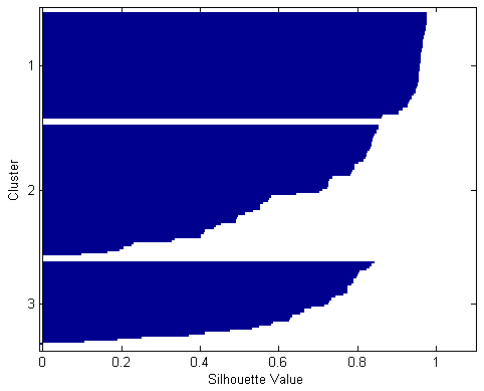


4. měření

Počet iterací: 9

Počáteční centroidy:
C1 = 5.1000 3.8000 1.5000 0.3000
C2 = 6.0000 3.4000 4.5000 1.6000
C3 = 5.9000 3.2000 4.8000 1.8000

Konečné centroidy:
C1 = 5.0060 3.4280 1.4620 0.2460
C2 = 5.8836 2.7410 4.3885 1.4344
C3 = 6.8538 3.0769 5.7154 2.0538

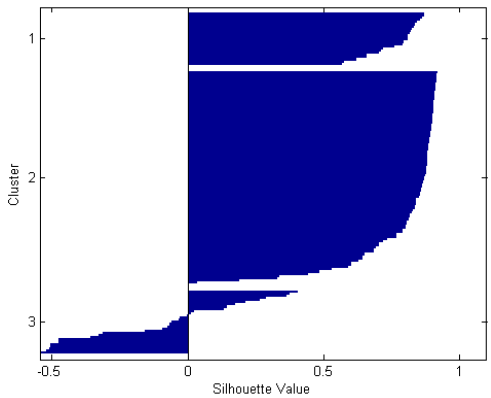


5. měření

Počet iterací: 7

Počáteční centroidy:
C1 = 4.9000 3.0000 1.4000 0.2000
C2 = 7.0000 3.2000 4.7000 1.4000
C3 = 4.8000 3.4000 1.9000 0.2000

Konečné centroidy:
C1 = 4.7333 3.1583 1.3917 0.2000
C2 = 6.3010 2.8866 4.9588 1.6959
C3 = 5.2310 3.5448 1.7000 0.3655



Zpracovala**: Radoslava Jandová**
Obor: **STM, Softwarové inženýrství**
E-mail: **jandora1@fel.cvut.cz**
Semestr: **zimní 2011/2012**

U opakovaných testů pro 3 shluky se již silueta nejeví jako dobrá. Ve 3 případech (měření 2., 4. a 5.) z 5 měření zasahuje silueta i do záporné části grafu. Zde jsou různé jak počáteční, tak i konečné hodnoty centroidů. U měření 1. a 3. se silueta jeví jako dobrá. V těchto případech jsou počáteční hodnoty odlišné, což odpovídá náhodnému výběru, ale konečné hodnoty jsou shodné.

Na základě těchto opakovaných měření lze konstatovat, že lépe vychází algoritmus pro 2 reprezentanty.

Zpracovala**: Radoslava Jandová** E-mail: **jandora1@fel.cvut.cz**
Obor: **STM, Softwarové inženýrství** Semestr: **zimní 2011/2012**