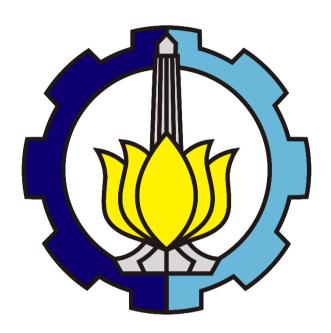
LAPORAN FINAL PROJECT TOPIK DALAM DATA MINING A



Disusun oleh:

Reza Presetya Prayogo 5116201009

Ari Effendi 5116201016

Addien Haniefardy 5116201049

Jurusan Teknik Informatika

Fakultas Teknologi Informasi
Institut Teknologi Sepuluh Nopember Surabaya
2017

1. Paper Acuan

"Using the Stability of Objects to Determine the Number of Clusters in Datasets"

2. Latar Belakang

Clustering merupakan suatu cara yang digunakan untuk membagi sejumlah data menjadi kelompok-kelompok data atau mengelompokkan suatu objek dengan objek yang lain berdasarkan similaritasnya. Semakin similar suatu objek/data dengan objek/data yang lain, semakin besar pula kemungkinan kedua objek tersebut tergabung dalam satu cluster (kelompok). Nilai similaritas dihitung menggunakan fitur-fitur/atribut-atribut yang terdapat pada setiap objek/data.

Dalam melakukan clustering, terdapat sejumlah metode yang bisa digunakan, diantaranya menggunakan metode K-Means. K-Means merupakan salah satu algoritma yang sering digunakan dalam klusterisasi karena mudah diimplementasikan dan relatif cepat. Huruf K pada K-Means menunjukkan jumlah kelompok yang ingin dibuat. Pada perkembangannya, nilai K menjadi nilai yang sangat penting untuk ditentukan dalam menentukan cluster yang paling optimal.

Untuk menentukan nilai K, banyak sekali pendekatan yang bisa dilakukan, diantaranya menggunakan algoritma Particle Swarm Optimization (PSO), Ant Colony Optimization (ACO), dan lain sebagainya. Pada paper ini, akan dikembangkan algoritma penentuan nilai K pada K-Means menggunakan nilai stabilitas objek. Nilai stabilitas objek sendiri dihitung menggunakan metode Silhouette.

3. Tujuan

Implementasi paper dalam melakukan optimasi nilai K pada K-Means menggunakan nilai stabilitas objek dengan metode Silhouette

4. Dataset

Pada final project kali ini, kami menggunakan dataset Zoo dengan 101 objek, 17 fitur, dan nilai K sama dengan 7 sebagai bentuk klaster optimal. Pada gambar pertama, bisa dilihat

```
1 (41) aardvark, antelope, bear, boar, buffalo, calf,
       cavy, cheetah, deer, dolphin, elephant,
       fruitbat, giraffe, girl, goat, gorilla, hamster,
       hare, leopard, lion, lynx, mink, mole, mongoose,
       opossum, oryx, platypus, polecat, pony,
       porpoise, puma, pussycat, raccoon, reindeer,
       seal, sealion, squirrel, vampire, vole, wallaby, wolf
2 (20) chicken, crow, dove, duck, flamingo, gull, hawk,
       kiwi, lark, ostrich, parakeet, penguin, pheasant,
       rhea, skimmer, skua, sparrow, swan, vulture, wren
3 (5) pitviper, seasnake, slowworm, tortoise, tuatara
4 (13) bass, carp, catfish, chub, dogfish, haddock,
       herring, pike, piranha, seahorse, sole, stingray, tuna
5 (4) frog, frog, newt, toad
6 (8) flea, gnat, honeybee, housefly, ladybird, moth, termite, wasp
7 (10) clam, crab, crayfish, lobster, octopus,
       scorpion, seawasp, slug, starfish, worm
```

101 objek yang tersebar ke dalam 7 klaster. Pada gambar kedua, bisa dilihat fitur (2-18) yang dimiliki oleh setiap objek beserta tipe nilainya.

```
Attribute Information: (name of attribute and type of value domain)

    animal name:

                    Unique for each instance
hair
                    Boolean
feathers
                    Boolean
eggs
                   Boolean
5. milk
                   Boolean
airborne
                   Boolean
aquatic
                   Boolean
predator
                   Boolean
toothed
                   Boolean
10. backbone
                    Boolean
11. breathes
                    Boolean
12. venomous
                    Boolean
                    Boolean
13. fins
14. legs
                   Numeric (set of values: {0,2,4,5,6,8})
15. tail
                    Boolean
domestic
                    Boolean
                    Boolean
17. catsize
                    Numeric (integer values in range [1,7])
18. type
```

Gambar 2 Fitur-fitur pada dataset Zoo beserta tipe nilainya

5. Mekanisme Algoritma

Sebelum masuk pada algoritma, terlebih dahulu kita definikan nilai R, K_{min} dan K_{max} . R merupakan jumlah partisi yang ingin dibuat. Jumlah partisi ini akan menentukan nilai stabilitas dari suatu objek. K_{min} merupakan nilai K minimal dan K_{max} merupakan nilai K maksimal yang diinginkan. Solusi nilai K yang didapatkan akan berada pada rentan $K_{min} \le K \le K_{max}$. Pseudocode algoritma ditunjukkan pada gambar xx.

```
For K = K_{min} to K_{max} do
         For r = 1 to R do
                  Generate a random starting partition (RP<sub>r</sub>) of objects of \Omega into K non-
empty classes
                  Execute K-Means using RP<sub>r</sub> as input to get partition Pr
         End
         For i = 1 to N do
                  Compute the singleton support PS<sub>i</sub>
                  For j = i + 1 to N do
                           Compute the pairwise support PS<sub>ij</sub>
                  End
         End
         For i = 1 to N do
                  Compute the individual stability index ST<sub>approx</sub>(i,K)
         Compute the global stability index, ST_{global}(K) for the case of K classes
End
Find the maximum value of ST<sub>global</sub>(K) over all tested of K
The optimal number of classes, K<sub>opt</sub>, will correspond to the maximum of ST<sub>global</sub>(K)
```

Untuk penjelasan lebih detail, kami berikan contoh dataset Zoo (N = 101) dengan nilai R sama dengan 1000, K_{min} sama dengan 2, dan K_{max} sama dengan 10. Tujuan dari algoritma

ini adalah untuk mendapatkan nilai K yang paling optimal (K_{opt}). Pada perulangan pertama (K = 2), dilakukan sub perulangan pertama untuk membuat R partisi. Jadi, dihasilkan 1000 partisi dengan setiap partisi terdiri dari 101 objek. Setelah itu, setiap partisi kemudian dilakukan pengklasteran menggunakan K-Means dengan nilai K sama dengan 2. Kemudian pada sub perulangan kedua, dilakukan 3 penghitungan. Pertama, penghitungan nilai Silhouette setiap objek pada semua partisi. Kedua, penghitungan nilai PS_i menggunakan persamaan 1. Ketiga, penghitungan nilai PS_{ij} menggunakan persamaan 2.

$$PS_{i} = \frac{\sum_{r=1}^{R} S_{r,i}}{\sum_{r=1}^{R} S_{r}}$$
 (1)

$$PS_{ij} = \frac{\sum_{r=1}^{R} S_{r,ij}}{\sum_{r=1}^{R} S_r}$$
 (2)

Dengan PS_i merupakan nilai stabilitas objek ke-i pada semua partisi. Dan PS_{ij} merupakan nilai stabilitas objek ke-i terhadap objek ke-j pada semua partisi. Lalu S_r merupakan nilai rata-rata dari penjumlahan nilai Silhouette semua objek pada satu partisi. Nilai PS_i dan PS_{ij} ini kemudian akan digunakan pada sub perulangan ketiga untuk menghitung nilai ST_{approx}(i,K) menggunakan persamaan 3.

$$ST_{approx}(i) = \frac{1}{N} \sum_{j=1(j\neq i)}^{N} \max\left(\frac{K}{K-1} * \left(PS_{ij} - \frac{1}{K}\right); K * \left(\frac{1}{K} - PS_{ij}\right)\right) + \frac{1}{N} \max\left(\frac{K^{N-1}}{K^{N-1} - (K-1)^{N-1}} * \left(PS_{i} - \frac{(K-1)^{N-1}}{K^{N-1}}\right); \frac{K^{N-1}}{(K-1)^{N-1}} * \left(\frac{(K-1)^{N-1}}{K^{N-1}} - PS_{i}\right)\right)$$

$$(3)$$

Setelah nilai ST_{approx}(i,K) didapatkan, dilakukan penghitungan ST_{global}(K) menggunakan persamaan 4.

$$ST_{global} = \frac{1}{N} \sum_{i=1}^{N} ST(i)$$
 (4)

Untuk perulangan kedua (K = 3), perulangan ketiga (K = 4), dan seterusnya lakukan hal yang sama sampai mendapatkan nilai $ST_{global}(K)$. Setelah kita memperoleh nilai ST_{global} untuk setiap K, langkah selanjutnya adalah mendapatkan nilai ST_{global} tertinggi dari nilai-nilai ST_{global} tersebut. K dengan nilai ST_{global} tertinggi merupakan K_{opt} .

6. Uji Coba

1. Mendapatkan partisi P_r menggunakan K-means

```
load('fdata.mat');
 kmin=2; %Jumlah K minimal
 kmax=4; %Jumlah K maksimal
 [N, Nfitur] = size(fdata);
 N=length(fdata); %Jumlah objek
 R=3; %Random start K-means
for k=kmin:kmax
     for r=1:R
          temp_kmeans = kmeans(fdata,k);
          for RPr=1:N
              for SRcount=1:N
                  for NfiturCount=1:Nfitur
                      Pr{1,k}{RPr,r}(SRcount,NfiturCount) = fdata(SRcount,NfiturCount);
              end
              for kmeanscount=1:N
                  Pr{1,k}{RPr,r}(kmeanscount,Nfitur+1) = temp_kmeans(kmeanscount,1);
              for silhouetteCount=1:N
                  \texttt{temp\_silhouette} = \texttt{silhouette}(\texttt{Pr\{1,k\}\{1,r\}(:,1:Nfitur), Pr\{1,k\}\{1,r\}(:,Nfitur+1),'SqEuclidean')};
              end
              for count_tmp_silh=1:N
                  Pr{1,k}{RPr,r}(count tmp silh,Nfitur+2) = temp silhouette(count tmp silh,1);
          end
     end
                                                                                     Go to Settings to activate Windows.
end
```

Gambar 6.1 Partisi Pr

2. a. Mendapatkan singleton support PS_i menggunakan Eq. (2)

Gambar 6.2.a Partition Score SRi

```
kmin=2;
 kmax=4;
 R=3;
 [N,Nfitur] = size(fdata);
 N=length(fdata); %Jumlah objek
for k=kmin:kmax
     SR_total = 0;
     for r=1:R
         SR tmp = 0;
         for count tmp SR=1:N
             SR_tmp = SR_tmp+Pr{1,k}{1,r}(count_tmp_SR,Nfitur+2);
         end
         SR(r,k) = SR tmp/N;
         SR_total = SR_total+SR_tmp/N;
     SR sum(1,k) = SR total;
 end
```

Gambar 6.2.a Partition Score Sr

Gambar 6.2.a singleton support PSi

b. Mendapatkan pairwise support PS_{ii} menggunakan Eq. (1)

```
kmin=2;
  kmax=4;
  R=3;
  [N,Nfitur] = size(fdata);
   N=length(fdata);
for k=kmin:kmax
for r=1:R
for i=1:
             for i=1:N+1
                     for j=i+1:N
                            if(Pr\{1,k\}\{1,r\}(i,Nfitur+1) == Pr\{1,k\}\{1,r\}(j,Nfitur+1))
                                 \underline{\mathtt{SRij}}\{\mathtt{r},\mathtt{k}\}\,(\mathtt{i},\mathtt{j})\ =\ \mathtt{SR}\,(\mathtt{r},\mathtt{k})\,;
                                  SRij\{r,k\}(i,j) = 0;
                           end
                     end
               end
         end
   end
```

Gambar 6.2.b partition score SRij

Gambar 6.2.b pairwise support PSij

3. Mendapatkan individual stability index ST(i,K) atau approximate variant ST_{approx}(i,K) menggunakan Eq. (9) dan (10).

```
kmin=2;
 kmax=4;
 R=3:
 [N, Nfitur] = size(fdata);
 N=length(fdata):
for k=kmin:kmax
     for i=1:N
          hasill = 0;
          hasi12 = 0;
          for j=i+1:N
              j1 = (k/(k-1))*(PSij\{1,k\}(i,j)-(1/k));
j2 = k*((1/k)-PSij\{1,k\}(i,j));
              if (j1 > j2)
                  hasill = hasill + jl;
               else
                   hasill = hasill + j2;
              end
          il = (k.^(N-1)/(k.^(N-1)-(k-1).^(N-1)))*(PSi(i,k)-((k-1).^(N-1)/k.^(N-1)));
           12 = (k.^{(N-1)/(k-1).^{(N-1)})*(((k-1).^{(N-1)/k.^{(N-1)})-PSi(i,k)); 
          if (i1 > i2)
              hasi12 = i1;
          else
              hasi12 = i2:
          end
          STapprox(i,k) = (hasill + hasil2)/N;
```

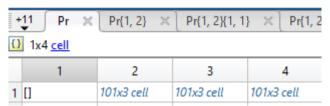
Gambar 6.3 approximate variant ST_{approx}

4. Mendapatkan global stability index menggunakan Eq. (12).

```
kmin=2;
kmax=4;
R=3;
[N,Nfitur] = size(fdata);
N=length(fdata);
for k=kmin:kmax
    tmp_STglobal = 0;
for i=1:N
    tmp_STglobal = tmp_STglobal + STapprox(i,k);
end
STglobal(1,k) = tmp_STglobal / N;
end
```

Gambar 6.4 global stability index ST_{global}

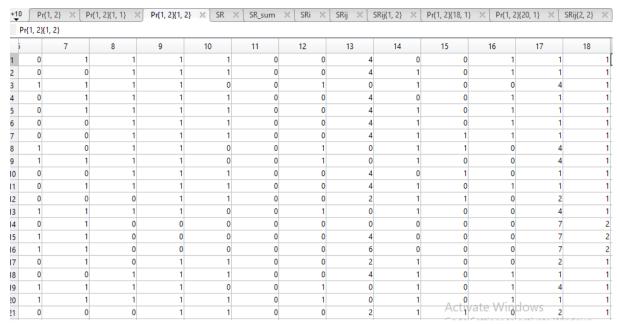
7. Hasil



Gambar 7.1 partisi kmin=2, kmax=4

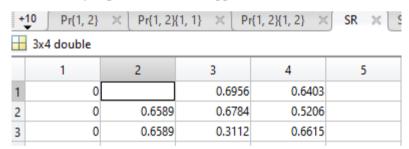


Gambar 7.2 random start R=3

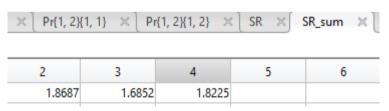


Gambar 7.3 partisi Pr menggunakan Kmeans

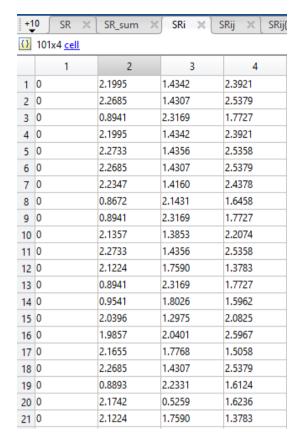
Kolom ke-1 sampai 17 merupakan fitur dari Dataset, sedangkan kolom ke-18 merupakan index cluster yang terbentuk menggunakan metode kmeans



Gambar 7.4 partition score Sr



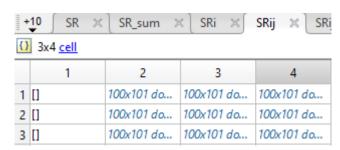
Gambar 7.5 sum partition score Sr

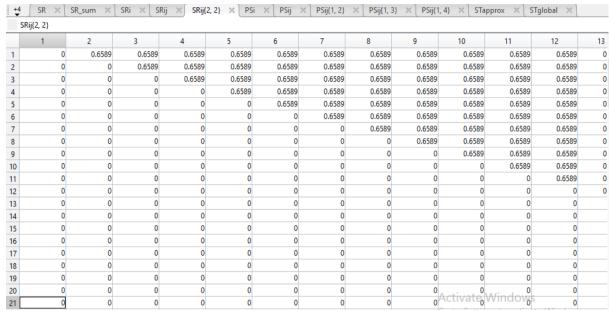


Gambar 7.6 partition score Sr,i

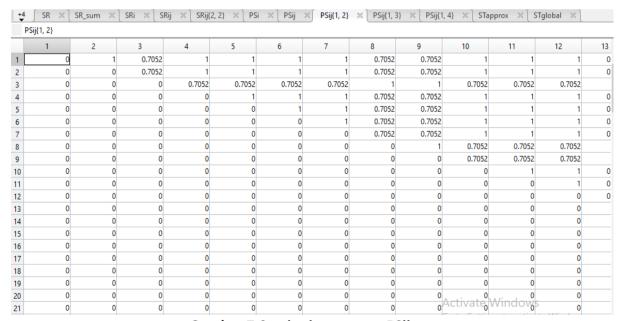
	1	2	3	4
1	0	3.9928	2.0619	3.7359
2	0	4.1179	2.0569	3.9636
3	0	1.6230	3.3310	2.7686
4	0	3.9928	2.0619	3.7359
5	0	4.1268	2.0639	3.9604
6	0	4.1179	2.0569	3.9636
7	0	4.0566	2.0358	3.8074
8	0	1.5743	3.0812	2.5704
9	0	1.6230	3.3310	2.7686
10	0	3.8769	1.9917	3.4475
11	0	4.1268	2.0639	3.9604
12	0	3.8528	2.5290	2.1527
13	0	1.6230	3.3310	2.7686
14	0	1.7321	2.5916	2.4930
15	0	3.7024	1.8653	3.2525
16	0	3.6046	2.9331	4.0555
17	0	3.9310	2.5544	2.3517
18	0	4.1179	2.0569	3.9636
19	0	1.6143	3.2105	2.5182
20	0	3.9469	0.7561	2.5358
21	0	3.8528	2.5290	2.1527

Gambar 7.7 singleton support PSi

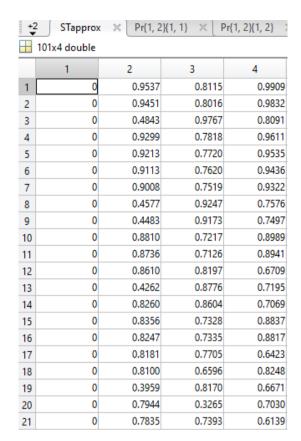




Gambar 7.8 partition score Srij



Gambar 7.9 pairwise support PSij



Gambar 7.10 approximate variant ST_{approx}(i,K)



Gambar 7.11 global stability index ST_{global}