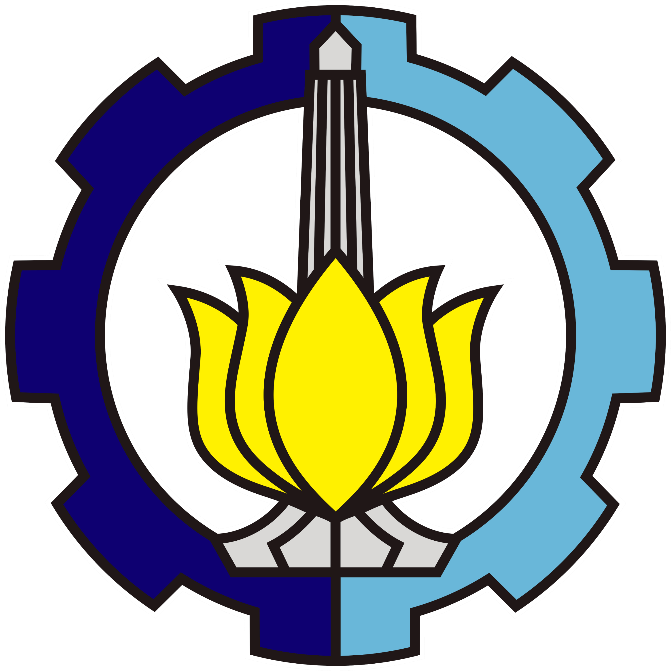
**LAPORAN FINAL PROJECT**

**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**”

1. 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.

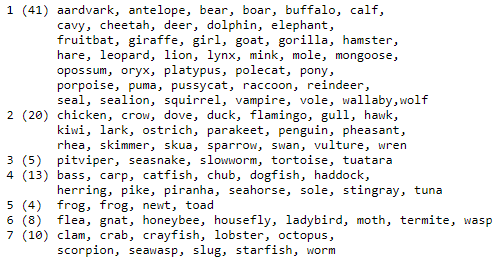
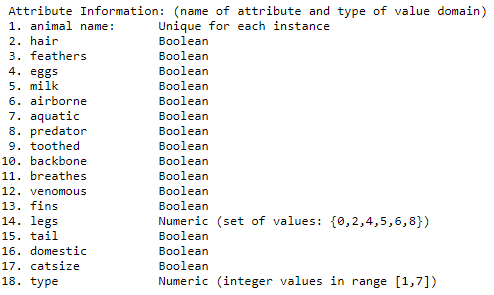
Dalam melakukan clustering, terdapat sejumlah metode yang bisa digunakan, diantaranya menggunakan metode K-Means. K-Means merupakan salah satu algoritma yang sering digunakan dalam klusterisasi karena mudah diimplementasikan dan relatif cepat. Huruf K pada K-Means menunjukkan jumlah kelompok yang ingin 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.

1. Tujuan

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

1. 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 101 objek yang tersebar ke dalam 7 klaster. Pada gambar kedua, bisa dilihat fitur (2-18) yang dimiliki oleh setiap objek beserta tipe nilainya.

Gambar 1 Dataset Zoo - 101 Objek yang terbagi ke dalam 7 klaster

Gambar 2 Fitur-fitur pada dataset Zoo beserta tipe nilainya

1. Mekanisme Algoritma

Sebelum masuk pada algoritma, terlebih dahulu kita definikan nilai R, Kmin dan Kmax. R merupakan jumlah partisi yang ingin dibuat. Jumlah partisi ini akan menentukan nilai stabilitas dari suatu objek. Kmin merupakan nilai K minimal dan Kmax merupakan nilai K maksimal yang diinginkan. Solusi nilai K yang didapatkan akan berada pada rentan Kmin ≤ K ≤ Kmax. Pseudocode algoritma ditunjukkan pada gambar xx.

**For** K = Kmin **to** Kmax **do**

**For** r = 1 **to** R **do**

Generate a random starting partition (RPr) of objects of Ω into K non-empty classes

Execute K-Means using RPr as input to get partition Pr

**End**

**For** i = 1 **to** N **do**

Compute the singleton support PSi

**For** j = i + 1 **to** N **do**

Compute the pairwise support PSij

**End**

**End**

**For** i = 1 **to** N **do**

Compute the individual stability index STapprox(i,K)

**End**

Compute the global stability index, STglobal(K) for the case of K classes

**End**

Find the maximum value of STglobal(K) over all tested of K

The optimal number of classes, Kopt, will correspond to the maximum of STglobal(K)

Untuk penjelasan lebih detail, kami berikan contoh dataset Zoo (N = 101) dengan nilai R sama dengan 1000, Kmin sama dengan 2, dan Kmax sama dengan 10. Tujuan dari algoritma ini adalah untuk mendapatkan nilai K yang paling optimal (Kopt). 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 PSi menggunakan persamaan 1. Ketiga, penghitungan nilai PSij menggunakan persamaan 2.

|  |  |
| --- | --- |
|  | (1) |
|  | (2) |

Dengan PSi merupakan nilai stabilitas objek ke-i pada semua partisi. Dan PSij merupakan nilai stabilitas objek ke-i terhadap objek ke-j pada semua partisi. Lalu Sr merupakan nilai rata-rata dari penjumlahan nilai Silhouette semua objek pada satu partisi. Nilai PSi dan PSij ini kemudian akan digunakan pada sub perulangan ketiga untuk menghitung nilai STapprox(i,K) menggunakan persamaan 3.

|  |  |
| --- | --- |
|  | (3) |

Setelah nilai STapprox(i,K) didapatkan, dilakukan penghitungan STglobal(K) menggunakan persamaan 4.

|  |  |
| --- | --- |
|  | (4) |

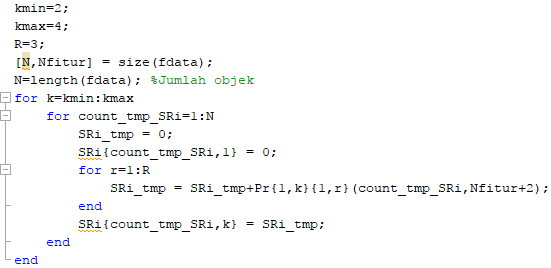
Untuk perulangan kedua (K = 3), perulangan ketiga (K = 4), dan seterusnya lakukan hal yang sama sampai mendapatkan nilai STglobal(K). Setelah kita memperoleh nilai STglobal untuk setiap K, langkah selanjutnya adalah mendapatkan nilai STglobal tertinggi dari nilai-nilai STglobal tersebut. K dengan nilai STglobal tertinggi merupakan Kopt.

1. Uji Coba
2. Mendapatkan partisi Pr menggunakan K-means

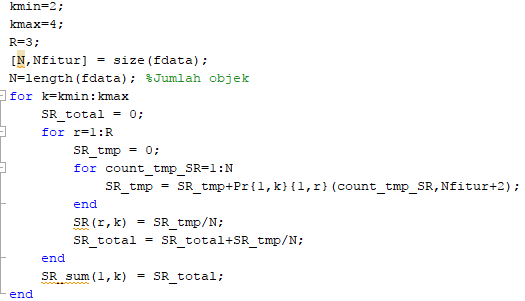


**Gambar 6.1 Partisi Pr**

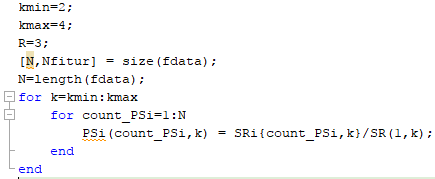
1. a. Mendapatkan singleton support PSi menggunakan Eq. (2)



**Gambar 6.2.a Partition Score SRi**

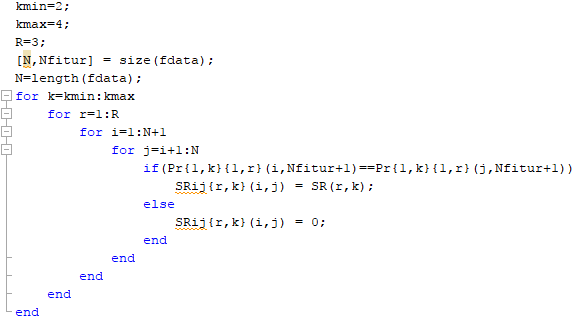


**Gambar 6.2.aPartition Score Sr**

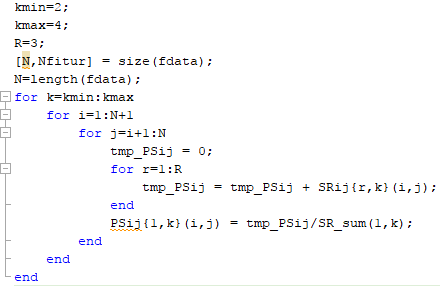


**Gambar 6.2.a singleton support PSi**

b. Mendapatkan pairwise support PSij menggunakan Eq. (1)

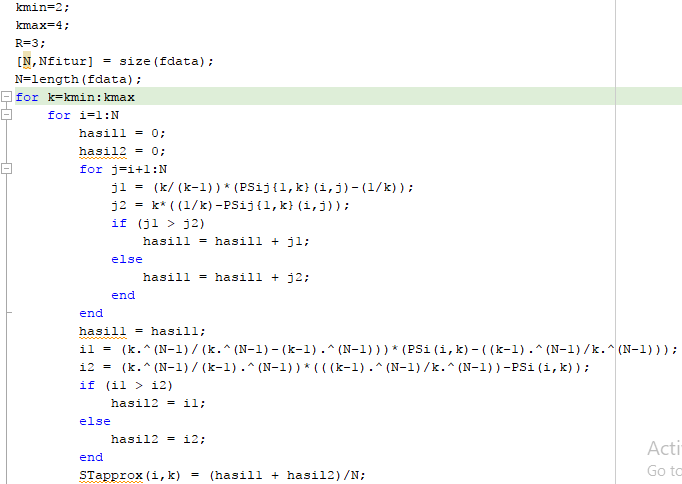


**Gambar 6.2.b partition score SRij**



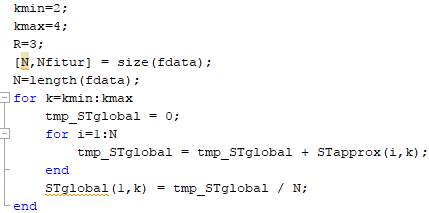
**Gambar 6.2.b pairwise support PSij**

1. Mendapatkan individual stability index ST(i,K) atau approximate variant STapprox(i,K) menggunakan Eq. (9) dan (10).



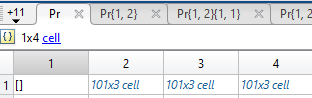
**Gambar 6.3 approximate variant STapprox**

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



**Gambar 6.4 global stability index STglobal**

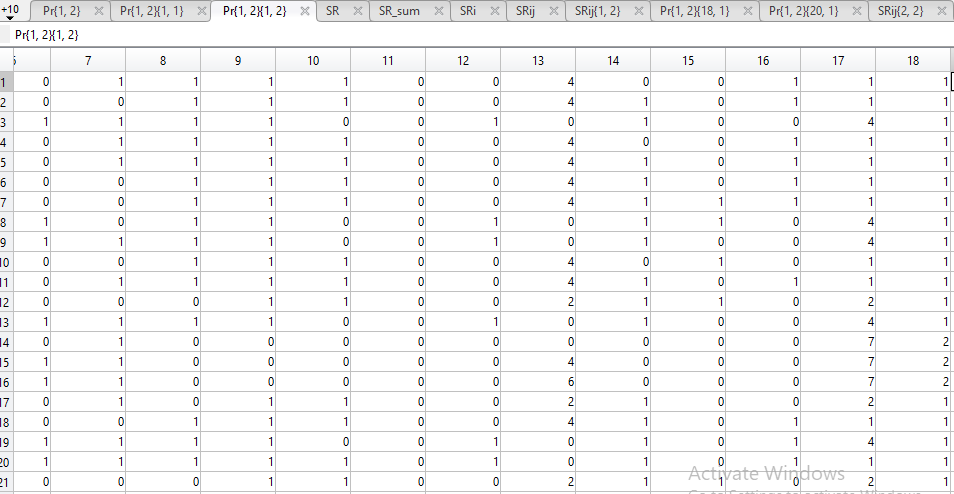
1. Hasil



**Gambar 7.1 partisi kmin=2, kmax=4**

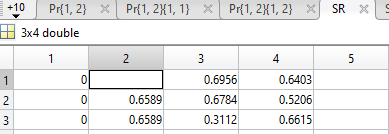


**Gambar 7.2 random start R=3**

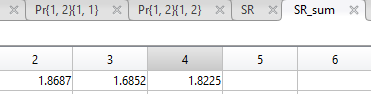


**Gambar 7.3 partisi Pr menggunakan Kmeans**

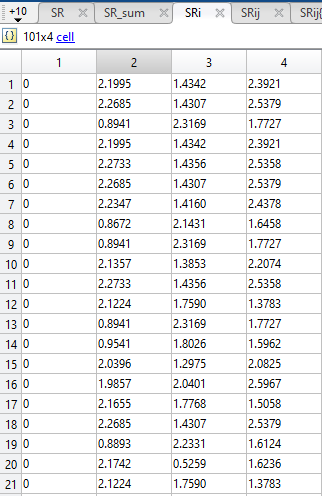
Kolom ke-1 sampai 17 merupakan fitur dari Dataset, sedangkan kolom ke-18 merupakan index cluster yang terbentuk menggunakan metode kmeans



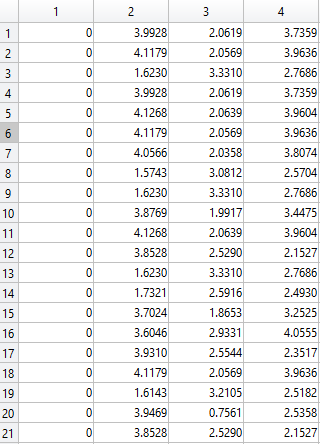
**Gambar 7.4 partition score Sr**



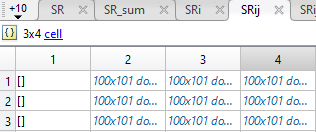
**Gambar 7.5 sum partition score Sr**

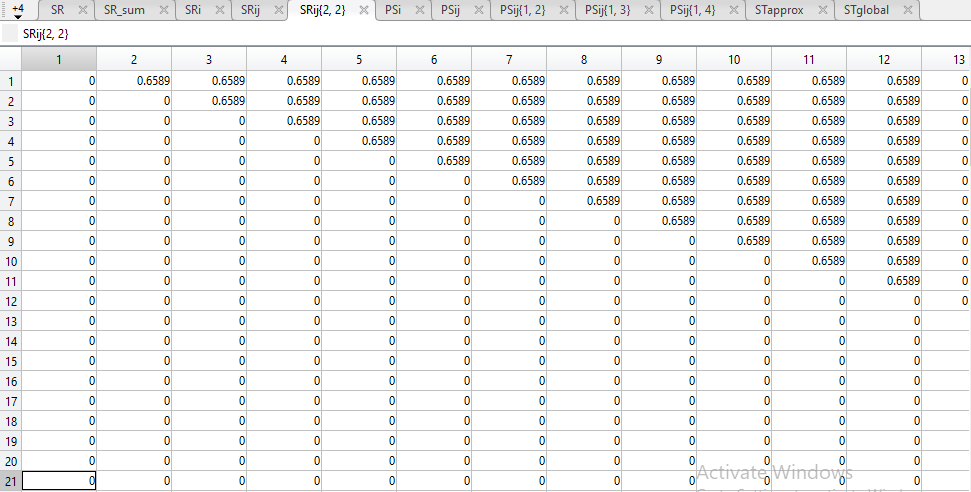


**Gambar 7.6 partition score Sr,i**

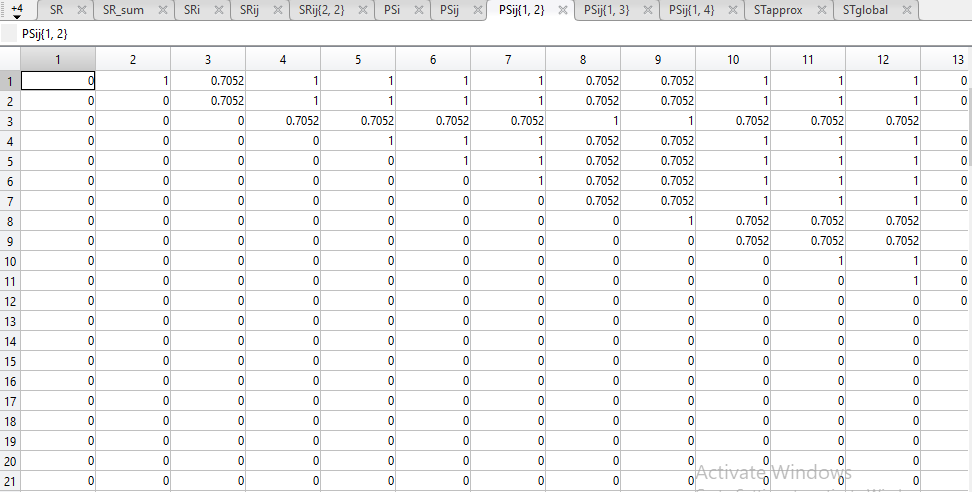


**Gambar 7.7 singleton support PSi**

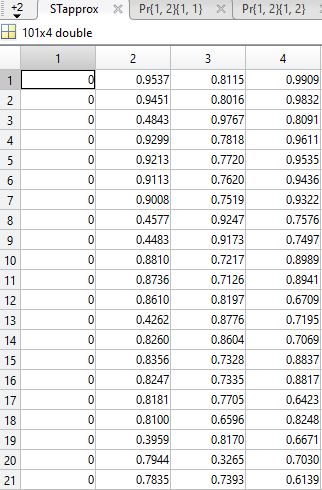




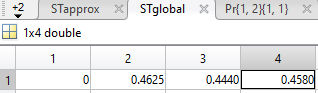
**Gambar 7.8 partition score Srij**



**Gambar 7.9 pairwise support PSij**



**Gambar 7.10 approximate variant STapprox(i,K)**



**Gambar 7.11 global stability index STglobal**