BIRCH – Balanced Iterative Reducing & Clustering with Hierarchies

Cluster the following data points using BIRCH.

B = Branching factor = 3

Threshold T = 5.

Points 2, 10, 12, 4, 25, 3, 30, 20, 11

→

Clustering feature CF = < n, LS, SS >

n = number of points in cluster

LS = linear sum of n points, $\sum_{i=1}^{n} x_i$

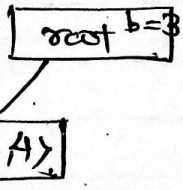
SS = square sum of n points, $\sum_{i=1}^{n} x_i^2$.

cluster centroid $x_0 = \frac{LS}{n}$.

cluster diameter $D = \sqrt{\frac{2n\,SS - 2\,LS^2}{n(n-1)}}$.

Step1 pt. 2. CF-Tree

CF = < 1, 2, 4 >

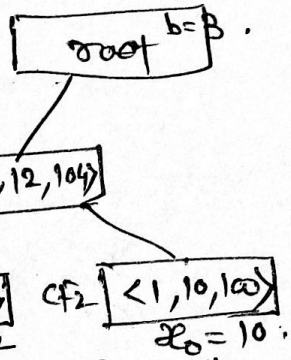


Step2 pt. 10 Try adding pt. 10 to CF_1. find new CF & diameter

$CF_1' = < 2, 2+10, 4+100 > = < 2, 12, 104 >$.

$d = \sqrt{\frac{2\times2\times104 - 2\times12\times12}{2(1)}} = \sqrt{60} = 8.$

d > T, Split CF.



Step3. pt. 12 Try adding pt. 12 to CF1 & CF2 find new CF & d.

CF1', pt. 12 is more closer CF2 as |12-10| < |12-2|

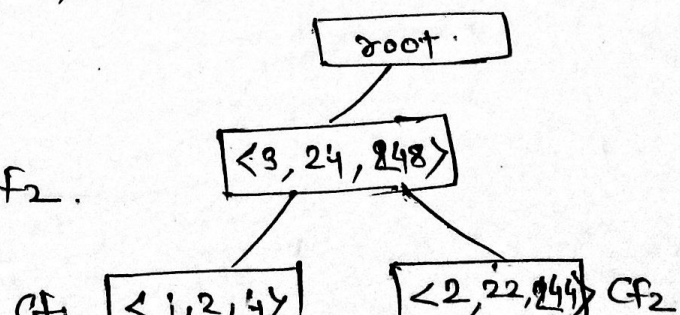
2 < 10.

Try adding pt. 12 to CF2.

$CF_2' = < 2, 10+12, 100+144 > = < 2, 22, 244 >$.

$d = \sqrt{\frac{2\times2\times244 - 2\times22\times22}{2(1)}} = 2$

d < 2 < T, add pt. 12 to CF2.



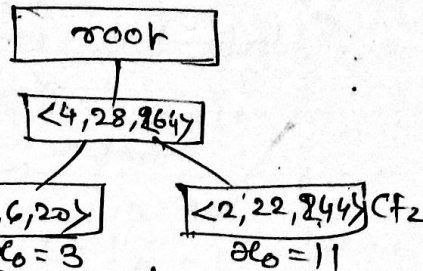
Step 4 – pr. 4 → pr. 4 is more closer to cf_1 as $|4-2| < |4-11|$.
$2 < 7$

Try adding pr. 4 to cf_1.

$cf_1' = \langle 2, 6, 20 \rangle$.

$$d = \sqrt{\frac{2 \times 2 \times 20 - 2 \times 6 \times 6}{2(1)}} = 2$$

$d < 2 < T$, add pr. 4 to cf_1.



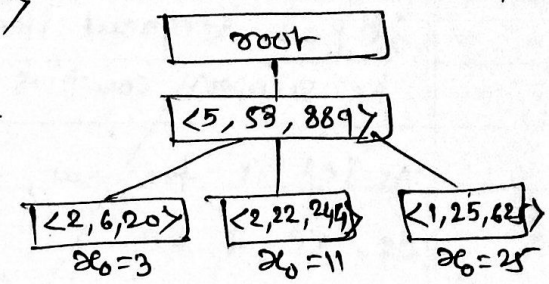
Step 5 – pr. 25 → pr 25 is more closer to cf_2 as $|25-3| > |25-11|$
$22 \qquad 14$

Try adding pr. 25 to cf_2

$cf_2' = \langle 3, 47, 769 \rangle$, $\langle 3, 47, 869 \rangle$

$$d = \sqrt{\frac{2 \times 2 \times 869 - 2 \times 47 \times 47}{2(1)}} = \text{-ve}$$

Hence add this as separate cf



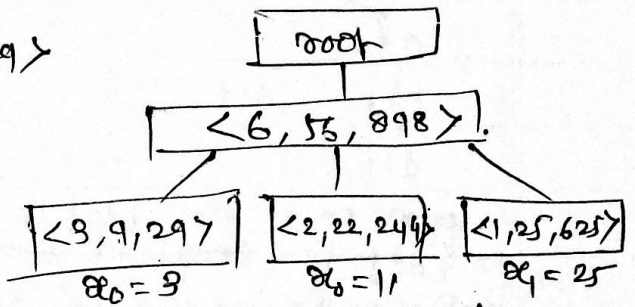
Step 6 – pr. 3 – pr is more closer to cf_1 as $|3-3| < |3-11| < |3-25|$
$0 \qquad 8 \qquad 22$

Try adding pr. 3 to cf_1.

$cf_1' = \langle 3, 6+3, 20+9 \rangle = \langle 3, 9, 29 \rangle$

$$d = \sqrt{\frac{2 \times 3 \times 29 - 2 \times 9 \times 9}{3(2)}} = 4.8$$

$d < T$, add pr. 3 to cf_1.



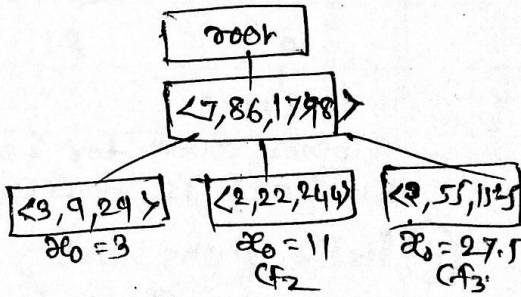
Step 7 – pr. 30 – pr 30 is more closer to cf_3 as $|30-25| < |30-11| < |30-3|$
$5 \qquad 19 \qquad 27$

Try adding pr. 30 to cf_3.

$cf_3' = \langle 2, 25+30, 625+900 \rangle = \langle 2, 55, 1525 \rangle$

$$d = \sqrt{\frac{2 \times 2 \times 1525 - 2 \times 55 \times 55}{2(1)}} = 5.$$

$d \leq T$, add pr. 30 to cf_3.



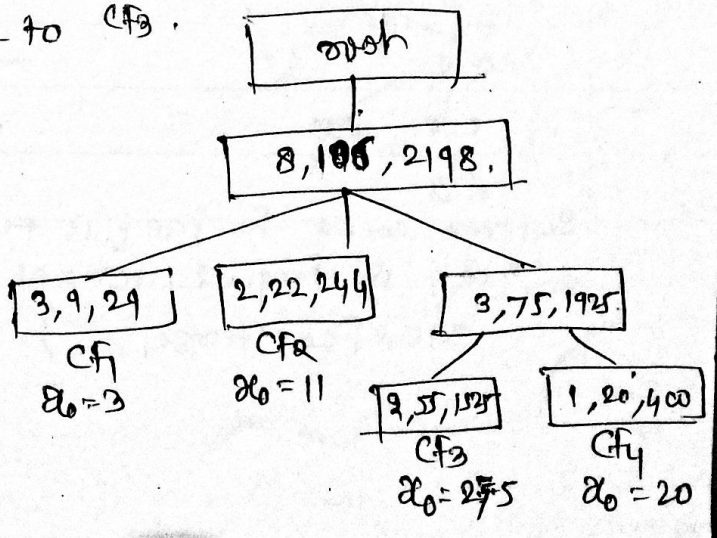
Step 8 – pr. 20 → pr 20 is more closer to cf_3.

Try adding to cf_3.

$cf_3' = \langle 3, 55+20, 1525+400 \rangle$
$= \langle 3, 75, 1925 \rangle$.

$$d = \sqrt{\frac{2 \times 3 \times 1925 - 2 \times 75 \times 75}{3(2)}} = 7.07$$

$d > T$, split cf_3' into 2 nodes.



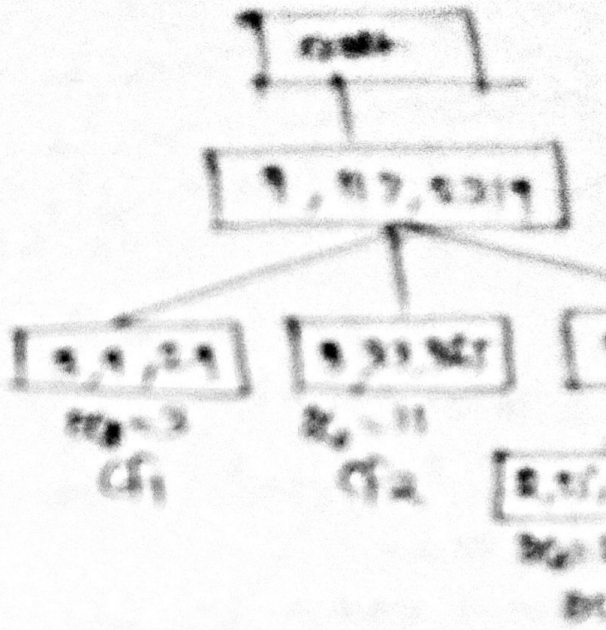
3- [illegible] is more close to C_2.
Adding [illegible] to C_2.

$C_2' = \langle 9, 11+11, 36+11 \rangle$

$= \langle 9, 22, 369 \rangle$.

$d = \sqrt{\frac{8\times9\times47 - 4\times33\times33}{9(8)}} = 1.41$.

$d < 7$, add [illegible] to C_2.



-final clusters are.

$C_1 = \langle 3, 4, 3 \rangle$

$C_2 = \langle 10, 12, 11 \rangle$

$C_3 = \langle 25, 20 \rangle$

$C_4 = \langle 20 \rangle$.

DBSCAN.

Use DBSCAN to cluster data points 2, 10, 12, 4, 25, 3, 30, 20, 11 with Eps radius $\epsilon = 2$ & min-pts = 2. Mark points clearly as core & noise points. If $\epsilon = 5$ is used, whether number of clusters are changed?

→ Step 0 Unvisited = {2, 10, 12, 4, 25, 3, 30, 20, 11}
Visited = { }
Noise = { }.

Step 1 - Compute ϵ neighbourhood & density of each point.

$N(2) = \{4, 3\}$	$d(2) = 2$	$N(3) = \{2, 4\}$	$d(3) = 3$
$N(10) = \{12, 11\}$	$d(10) = 2$	$N(30) = \{\}$	$d(30) = 1$
$N(12) = \{10, 11\}$	$d(12) = 2$	$N(20) = \{\}$	$d(20) = 1$
$N(4) = \{2, 3\}$	$d(4) = 2$	$N(11) = \{10, 12\}$	$d(11) = 2$
$N(25) = \{\}$	$d(25) = 1$		

Core-pts = {2, 10, 12, 4, 3, 11}.
Noise-pts = {25, 30, 20}.

Step 2 - Iterate over each core pt.

pt. 2 visited = {2}. $C_1 = \{2\}$.
$N(2) = \{4, 3\}$. As 4 is core pt.
p' = 4 visited = {2, 4} $N(2) = N(2)$ union $N(4)$.
$N(2) = \{4, 3\} \cup \{2, 3\} = \{2, 3, 4\}$. $C_1 = \{2, 4\}$.
p' = 3 visited = {2, 4, 3}.
As 3 is core pt.
$N(2) = N(2) + N(3) = \{2, 3, 4\} \cup \{2, 4\} = \{2, 3, 4\}$
$C_1 \in (2, 4, 3)$.
$C_1 = \{2, 4, 3\}$.

pt. 10 visited = {2, 4, 3, 10}.
$C_2 = \{10\}$ $N(10) = \{12, 11\}$.
p' = 12 visited = {2, 4, 3, 10, 12}
As 12 is core pt. $N(10) = N(10) \cup N(12) = \{12, 11\} \cup \{10, 11\} = \{12, 11, 10$
$C_2 = \{10, 12\}$.
p' = 11 visited = {2, 4, 3, 10, 12, 11}.
As 11 is core pt $N(10) = N(10) \cup N(11) = \{12, 11, 10\} \cup \{10, 12\} = \{12, 11, 10$
$C_2 = \{10, 12, 11\}$.
p' = 10 10 is already visited.
$C_2 = \{10, 12, 11\}$.

visited = {2,4,3,10,12,11}
noise = {25,30,20}.
Clusters C_1 = {2,4,3}
C_2 = {10,12,11}.

Step 6 $\epsilon = 5$, lets repeat the process.

N(2) = {4,3} d(2) = 3
N(12) = {10,11} d(12) = 3
N(25) = {30,20} d(25) = 3
N(3) = {2,4} d(3) = 3
N(20) = {25,30} d(20) = 3

N(10) = {12,11} d(10) = 3.
N(4) = {2,3} d(4) = 3
N(3) = {2,4} d(3) = 3
N(30) = {25,20} d(30) = 3.
N(11) = {10,12} d(11) = 3.

CorePb = {2,10,12,4,25,3,30,20,11}
Noise = { }.
Visited = { } Unvisited = { }.

Step 7 pt 2, visited = {2}. C_1 = {2}. as 2 is core pt form new cluster.
N(2) = {4,3}.
p' = 4 visited = {2,4} As 4 is core pt N(2) = N(2) ∪ N(4)
= {4,3} ∪ {2,3}
C_1 = {2,4} = {4,3,2}.
p' = 3 visited = {2,4,3} As 3 is core pt N(2) = N(2) ∪ N(3)
= {4,3,2} ∪ {2,4}
C_1 = {2,4,3} = {4,3,2}.
C_1 = {2,4,3} As all points in N(2) are visited now.

Step 8 pt. 10 visited = {2,4,3,10} C_2 = {10} as 10 is core pt.
N(10) = {12,11}
p' = 12 visited = {2,4,3,10,12}
As 12 is core pt N(10) = N(10) ∪ N(12) = {12,11} ∪ {10,11}
C_2 = {10,12}. = {12,11,10}.
p' = 11 visited = {2,4,3,10,12,11}.
As 11 is core pt, N(10) = N(10) N(11) = {12,11,10} ∪ {10,12}
= {12,11,10}.
C_2 = {10,12,11}. As all points in N(12) are visited now

Step 9 pt. 25 visited = {2,4,3,10,12,11,25} C_3 = {25} as 25 is core pt.
N(25) = {30,20}.
p' = 30 visited = {2,4,3,10,12,11,25,30}.
As 30 is core pt. N(25) = N(25) ∪ N(30) = {30,20} ∪ {25,20}
= {30,20,25}.
C_3 = {25,30}

p = 20 visited = {2, 4, 3, 10, 12, 11, 25, 30, 20}.

As 20 is core pt, $N(25) = N(25) \cup N(20) = \{30, 20, 25\} \cup \{25, 30\}$.

$= \{30, 20, 25\}$.

$C_3 = \{25, 30, 20\}$

As all points in N(25) are visited now.

All points are visited, now.

final clusters $C_1 = \{2, 4, 3\}$

$C_2 = \{10, 12, 11\}$.

$C_3 = \{25, 30, 20\}$.

As ϵ is changed from 2 to 5, number of clusters increased by one.

Evaluation of clustering – Hopkins stat.

finding out uniformity in data distribution.

If Hopkins stat = H ≤ 0.5, uniformly distributed.

otherwise not.

Eg. Determine weather following dataset has clustering tendency.

{2, 10, 12, 4, 25, 3, 30, 20, 11} = D.

→ Step 1 – Randomly sample 5 points from D.

Sample = {2, 12, 25, 30, 11}.

find out pts which are nearest to each pt in sample & sum all those nearest distances.

$\sum x_i = d(2,3) + d(12,11) + d(25,20) + d(30,25) + d(11,10)$,

$= 1 + 1 + 5 + 5 + 1$

$= 13$.

Step 2 – Randomly sample 5 points from but don't place them back in D.

$s_1 = \{12\}$. $D = \{2, 10, 4, 25, 3, 30, 20, 11\}$.

$s_2 = \{30\}$ $D = \{2, 10, 4, 25, 3, 20, 11\}$

$s_3 = \{4\}$ $D = \{2, 10, 25, 3, 20, 11\}$

$s_4 = \{11\}$ $D = \{2, 10, 25, 3, 20\}$

$s_5 = \{3\}$ $D = \{2, 10, 25, 20\}$

$y_1 = d(12,11) = 1$.

$y_2 = d(30,25) = 5$.

$y_3 = d(4,3) = 1$.

$y_4 = d(11,10) = 1$.

$y_5 = d(3,2) = 1$.

$\sum y_i = 1 + 5 + 1 + 1 + 1 = 9$.

$$H = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i}$$

$$= \frac{9}{13 + 9}$$

$$= \frac{9}{22}$$

$$= 0.4.$$

Looks data is uniformly distributed as per Hopkins stat so its not much suitable for data clustering.

Evaluation of Clustering – Elbow method

Determining number of clusters.

Q Consider following dataset, using different values of clusters. i.e. k = 2, 3, 4, 5, 6. determine best value of k using elbow techn.

→ D = {2, 10, 12, 4, 25, 3, 30, 20, 11}.

Assuming Manhattan dist^n as measure, & initial centroid 3, 10.

k=2, C_1 = {3, 2, 4}.
C_2 = {10, 12, 25, 30, 20, 11}

Var(k) = sum of squared dist^n of each pt from its centroid

=2

$= (3-3)^2 + (3-2)^2 + (3-4)^2 +$
$(10-10)^2 + (12-10)^2 + (25-10)^2 + (30-10)^2 + (20-10)^2 + (11-10)^2$

$= 0+1+1+0+4+225+400+100+1$

$= 732$

k=3 C_1 = {3, 2, 4}
C_1=3 C_2 = {10, 12, 11}
C_2=10 C_3 = {25, 30, 20}.
C_3=25

$Var(k=3) = (3-3)^2 + (3-2)^2 + (3-4)^2$
$+ (10-10)^2 + (12-10)^2 + (11-10)^2$
$+ (25-25)^2 + (25-30)^2 + (25-20)^2$

$= 0+1+1+0+4+1+0+25+25 = 57$

k=4
C_1=3 C_1 = {3, 2, 4}
C_2=10 C_2 = {10, 12, 11}
C_3=20 C_3 = {20, 25, }
C_4=30 C_4 = {30, }

$Var(k=4) = (3-3)^2 + (3-2)^2 + (3-4)^2$
$+ (10-10)^2 + (10-12)^2 + (10-11)^2$
$+ (20-20)^2 + (20-25)^2 +$
$+ (30-30)^2$

$= 0+1+1+0+4+1+0+25+0 = 32$

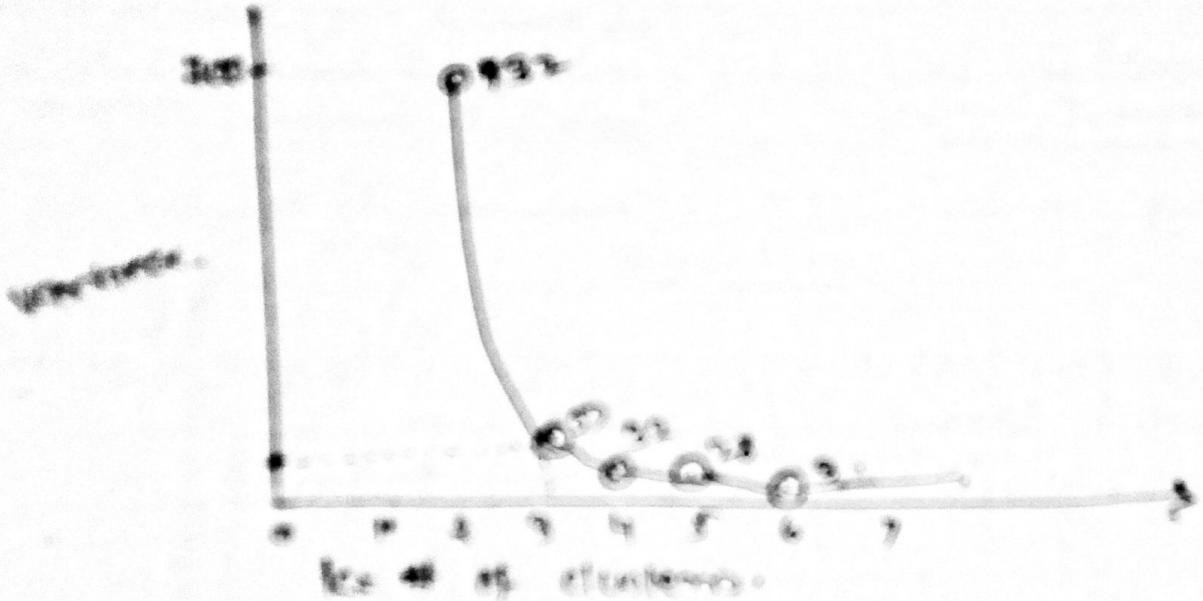
k=5. C_1 = {3, 2, 4}
C_1=3 C_2 = {10}
C_2=10 C_3 = {12, 11}
C_3=12 C_4 = {20}
C_4=20 C_5 = {25, 30}
C_5=25

$Var(k=5) = (3-3)^2 + (3-2)^2 + (3-4)^2$
$+ (10-10)^2$
$+ (12-12)^2 + (12-11)^2$
$+ (20-20)^2 + (25-25)^2 + (25-30)^2$

$= 0+1+1+0+0+1+0+0+25 = 28.$

k=6.
C_1=3 C_1 = {3, 2, 4}
C_2=10 C_2 = {10, 11}
C_3=12 C_3 = {12}
C_4=20 C_4 = {20}
C_5=25 C_5 = {25}
C_6=30. C_6 = {30}

$Var(k=6) = 0+1+1$
$+0+1$
$+0+0+0+0 = 3$



Here k=3, reduction in variance is not much more, hence will select maximum of clusters value i.e. k value = 3.

Q. Using data on distances given below, compute silhouette coefficient for each point, each of two clusters and the overall clustering. Cluster 1 contains $\{P_1, P_2\}$, cluster 2 contains $\{P_3, P_4\}$.

	P_1	P_2	P_3	P_4
P_1	0			
P_2	0.1	0		
P_3	0.65	0.70	0	
P_4	0.55	0.60	0.30	0

→ for P_1 $a = 0.1$ avg dist^n bet^n P_1 & all other points in C_1.

$$b = \frac{0.65 + 0.55}{2} = 0.6$$

= avg dist^n bet^n P_1 & all other points in C_2

$$s = \frac{b-a}{\max(a,b)} = \frac{0.6 - 0.1}{\max(0.1, 0.6)} = \frac{0.5}{0.6} = \boxed{0.833}$$

for P_2 $a = 0.1$

$$b = \frac{0.7 + 0.6}{2} = 0.65$$

$$s = \frac{b-a}{\max(a,b)} = \frac{0.65 - 0.1}{0.65} = \boxed{0.846}$$

for P_3 $a = 0.3$

$$b = \frac{0.65 + 0.7}{2} = 0.675$$

$$s = \frac{0.675 - 0.3}{0.675} = \boxed{0.555}$$

for P_4 $a = 0.3$

$$b = \frac{0.55 + 0.6}{2} = 0.575$$

$$s = \frac{0.575 - 0.3}{0.575} = \boxed{0.478}$$

for cluster C_1 $s = \frac{0.833 + 0.846}{2} = 0.839$

for cluster C_2 $s = \frac{0.555 + 0.478}{2} = 0.516$

for overall clustering $s = \frac{0.839 + 0.516}{2} = 0.677$

Q. How does F-score help in quantifying cluster quality? Given clustering results of newspaper articles dataset given below, compute F-score for cluster representing metro & financial articles.

Cluster	Entertain	fin	foreign	Metro	National	Sports	Total
#1	1	1	5	11	4	676	698
#2	27	89	333	827	253	33	1562
#3	126	465	8	105	16	29	749
Total	154	555	346	943	273	738	3009

$$f = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

→ for cluster 1

class Metro. $f = \dfrac{2 \times \frac{11}{676} \times \frac{11}{943}}{\frac{11}{676} + \frac{11}{943}} = 0.0135$

class financial $f = \dfrac{2 \times \frac{1}{676} \times \frac{1}{555}}{\frac{1}{676} + \frac{1}{555}} = 0.00162$

for cluster 2

class Metro $f = \dfrac{2 \times \frac{827}{1562} \times \frac{827}{943}}{\frac{827}{1562} + \frac{827}{943}} = 0.660$

class financial $f = \dfrac{2 \times \frac{89}{1562} \times \frac{89}{555}}{\frac{89}{1562} + \frac{89}{555}} = 0.084$

for cluster 3

class Metro $f = \dfrac{2 \times \frac{105}{749} \times \frac{105}{943}}{\frac{105}{749} + \frac{105}{943}} = 0.124$

class financial $f = \dfrac{2 \times \frac{465}{749} \times \frac{465}{555}}{\frac{465}{749} + \frac{465}{555}} = 0.713$