

Understanding and Enhancement of Internal Clustering Validation Measures

Batch No:1

P.Bhavana(Y12CS812)

D.Sai Tarun (Y12CS829)

G.Anvesh Babu(Y12CS848)

GUIDE:

Mr. P. Venkateswara Rao

Associate Professor

CSE Department

ABSTRACT

In general, clustering validation can be categorized into two classes, external clustering validation and internal clustering validation. We already have 11 widely used internal clustering validation measures for crisp clustering. But these existing measures have certain limitations in different application scenarios.

As an alternative choice, there is a new internal clustering validation measure, named clustering validation index based on nearest neighbors (CVNN), which is based on the notion of nearest neighbors. This measure can dynamically select multiple objects as representatives for different clusters in different situations.

Introduction

- ▶ Clustering validation, which evaluates the goodness of clustering results, has long been recognized as one of the vital issues essential to the success of clustering applications.

- ▶ Types of clustering validation:

External Clustering Validation

Internal Clustering Validation

- ▶ Why Internal clustering validation?

Internal Clustering Validation Measures

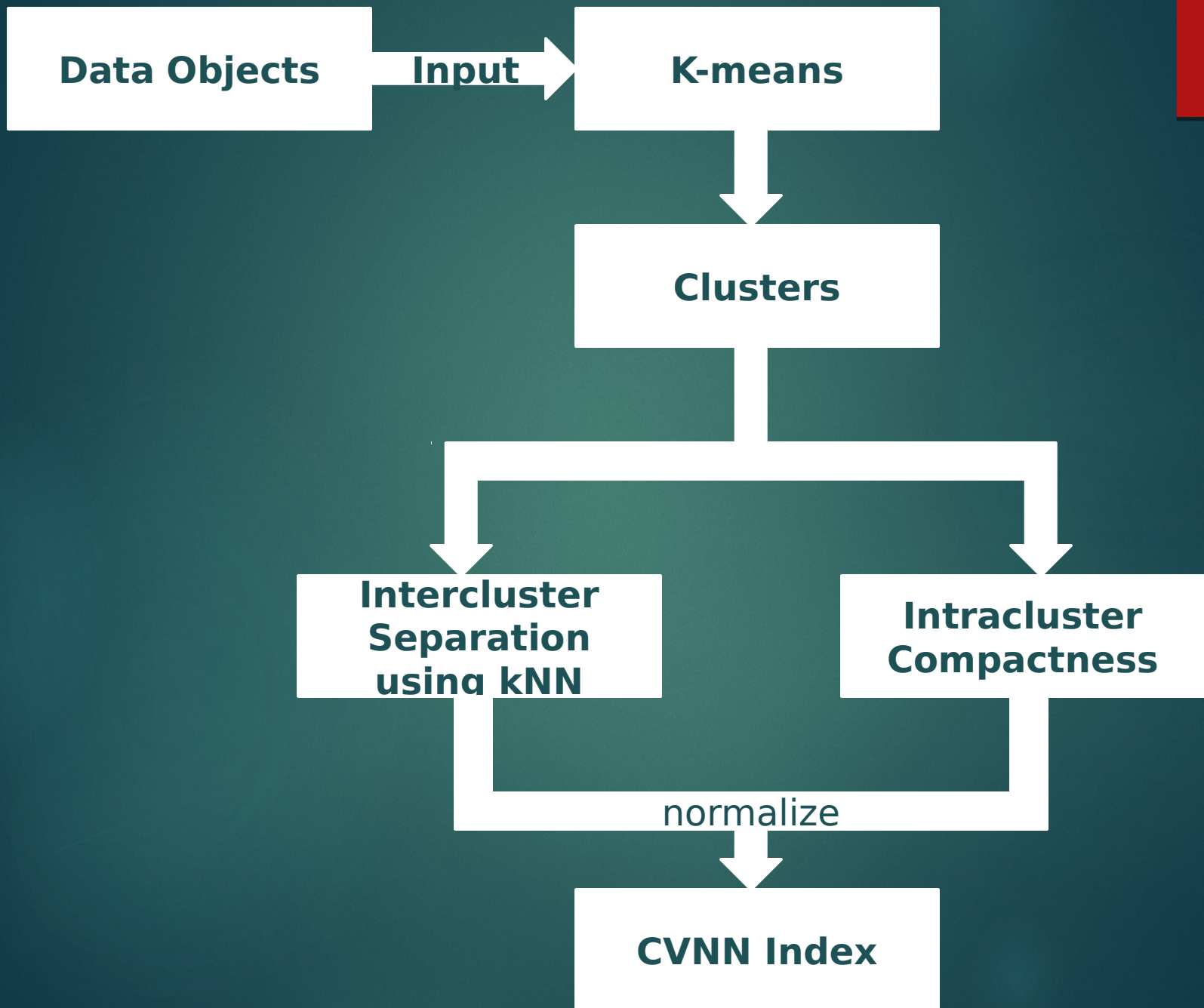
- ▶ The measures are based on two criteria:

Compactness:

It measures how closely related the objects in the cluster are.

Seperation:

It measures how distinct or well separated a cluster is from other clusters.



K-means algorithm:

- ▶ Randomly select 'c ' cluster centers.
- ▶ Calculate distance between each data object and cluster center.
- ▶ Assign the data object to the cluster center whose distance from the center is the minimum of all the cluster centers.
- ▶ Recalculate the cluster center using $v_{i=} (1/c_i) x_i$
where c_i represents the number of data objects in i^{th} cluster &
 x_i is the data object i .
- ▶ Recalculate the distance between each data object and new obtained cluster centers.
- ▶ If no data object was reassigned then stop, else repeat step 3.

CVNN Measure Calculation

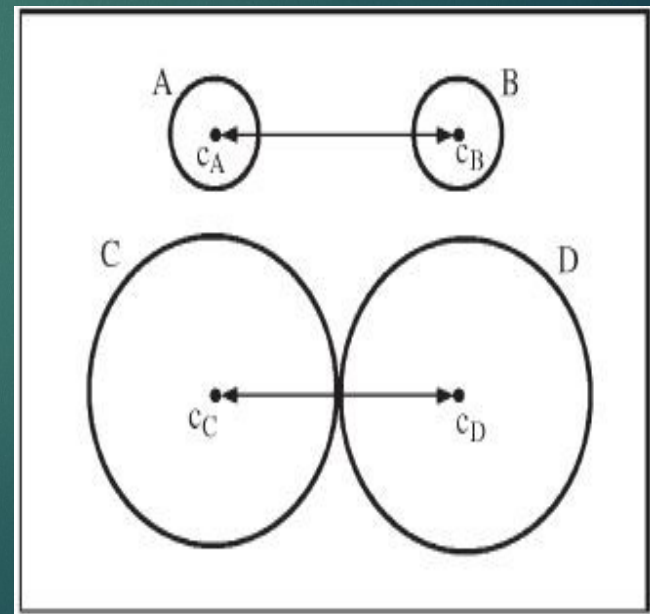
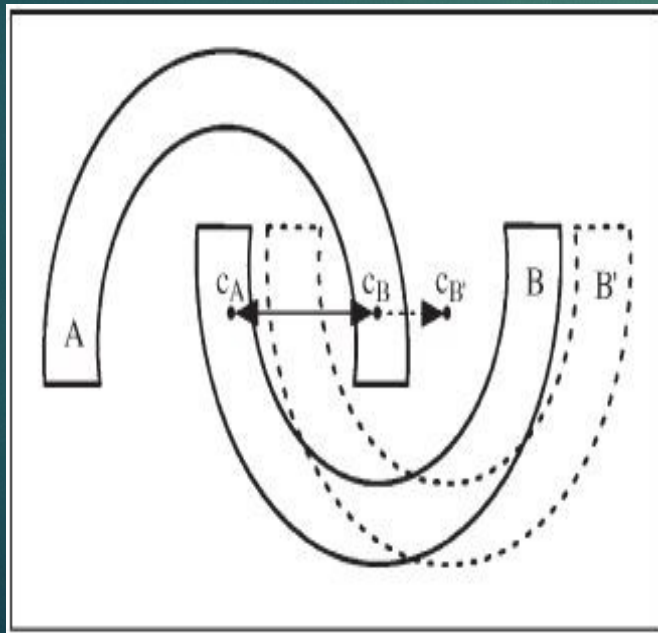
- ▶ Clustering Validation index based on Nearest Neighbors.
- ▶ A new Internal validation measure is proposed based on the notion of nearest neighbors.
- ▶ It is complementary to the existing measures.
- ▶ Based on


Intercluster Separation

Intraccluster Compactness

CVNN Index

- ▶ The intercluster separation is evaluated based on objects that carry the geometrical information of each cluster.
- ▶ Since one single object cannot reveal the geometrical information of the entire cluster, multiple objects are used as representatives.



- 
- ▶ For any data object in a cluster, its kNNs should also be in the same cluster.
 - ▶ If an object is located in the center of a cluster and is surrounded by objects in the same cluster, it is well separated from other clusters and thus contributes little to the intercluster separation.
 - ▶ If an object is located at the edge of a cluster and is surrounded mostly by objects in other clusters, it connects to other clusters tightly and thus contributes a lot to the intercluster separation.

We define the intercluster separation mainly in four steps

First, for each object in each cluster, find out whether at least one of its $kNNs$ is in other clusters.

Second, for objects with positive answers, assign a weight to each of them (q_i/k). These are the objects that best represent for the entire clusters; for objects with negative answers, the weight is zero.

Third, calculate the average weight of objects in the same cluster.

Finally, take the maximum average weight among all clusters as the intercluster separation.



$$\text{Sep}(\text{NC}, k) = \max_{i=1,2,\dots,\text{NC}} \left(\left(\frac{1}{n_i} \right) \sum_{j=1,2,\dots,n_i} (q_j/k) \right).$$

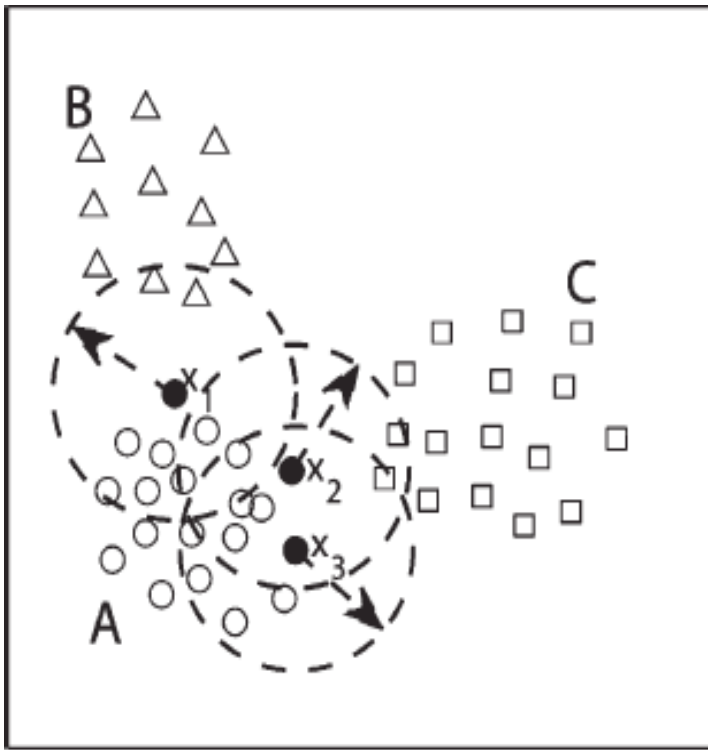
Where NC is the cluster number,

k is the number of nearest neighbors,

n_i is the number of objects in the i^{th} cluster C_i ,

O_j is the j^{th} object in C_i ,

and q_j is the number of nearest neighbors of O_j which are not in cluster C_i .



Example of intercluster Separation calculation

- Take $k=10$
- Objects x_1, x_2, x_3
- Objects in cluster – 20.
- $\text{Sep}_A(3,10)=?$

$$\text{Sep}(\text{NC}, k) = \max_{i=1,2,\dots,\text{NC}} \left(\left(\frac{1}{n_i} \right) \sum_{j=1,2,\dots,n_i} (q_j/k) \right).$$

Intraccluster compactness(Com)

► $\text{Com}(\text{NC}) = \sum_i [(2/n_i \cdot (n_i - 1)) \sum_{x, y \in C_i} d(x, y)].$

where NC is the cluster number,

n_i is the number of objects in the i^{th} cluster C_i ,

and x and y are two different objects in C_i .

- This measure is mainly based on the average pairwise distance between objects in the same cluster.
- In general, lower value of Com indicates a better intraccluster compactness.

CVNN Index:

- ▶ Based on the intercluster separation and intracluster compactness internal CVNN is defined.

- ▶ **CVNN Index:**

It is calculated in three steps

1. Calculate $Sep_{norm}(NC, k)$

$$Sep_{norm}(NC, k) = Sep(NC, k) / (\max_{NC_{min} \leq NC \leq NC_{max}} Sep(NC, k))$$

2. Calculate $Com_{norm}(NC)$

$$Com_{norm}(NC) = Com(NC) / (\max_{NC_{min} \leq NC \leq NC_{max}} Com(NC)).$$

- ▶ Finally

$$\mathbf{CVNN}(\mathbf{NC}, k) = \mathbf{Sep}_{\text{norm}}(\mathbf{NC}, k) + \mathbf{Com}_{\text{norm}}(\mathbf{NC}).$$

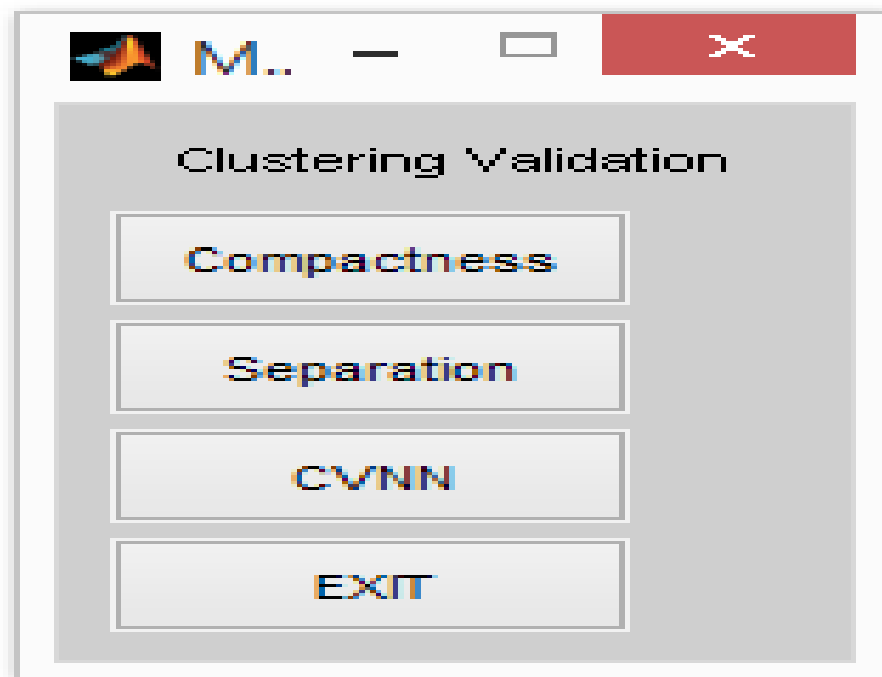
- ▶ Here normalize them to the same range before adding them up, since they should have the same order of magnitude.
- ▶ A lower value of CVNN indicates a better clustering result.

HARDWARE REQUIREMENTS:

- ▶ System : HP Pro 3330
- ▶ Hard Disk : 40 GB.
- ▶ Monitor : 18.5" LED Monitor.
- ▶ Ram : 1333 MHz.

SOFTWARE REQUIREMENTS:

- ▶ Operating system : Windows XP/7.
- ▶ Coding Language: MATLAB



File Edit View Graphics Debug Parallel Desktop Window

Current Folder: C:\Use

Shortcuts How to Add What's New

Variable Editor - count

Stack: Base No valid p

count <1x3 double>

	1	2	3	4	5
1	50	45	55		
2					
3					
4					
5					
6					
7					
8					
9					
10					


```
File Edit Debug Parallel Desktop Window Help
: [Icons] [Icons] [Icons] [Icons] [Icons] [Icons] [Icons] [Icons] [Icons] [Icons] Current Folder: C:\Users\bhavana\Desktop\paper\mainp
: Shortcuts [Icon] How to Add [Icon] What's New

Workspace
Variable Editor
enter number of clusters
3
enter initial cluster centre 20
enter initial cluster centre 70
enter initial cluster centre 140

Compact =

1.6731e+003

Compact =

1.8478e+003

Compact =

3.2959e+003

Sum =

6.8167e+003

fx >>
```

File Edit View Graphics Debug Parallel Desktop Window

Current Folder: C:\

Shortcuts How to Add What's New

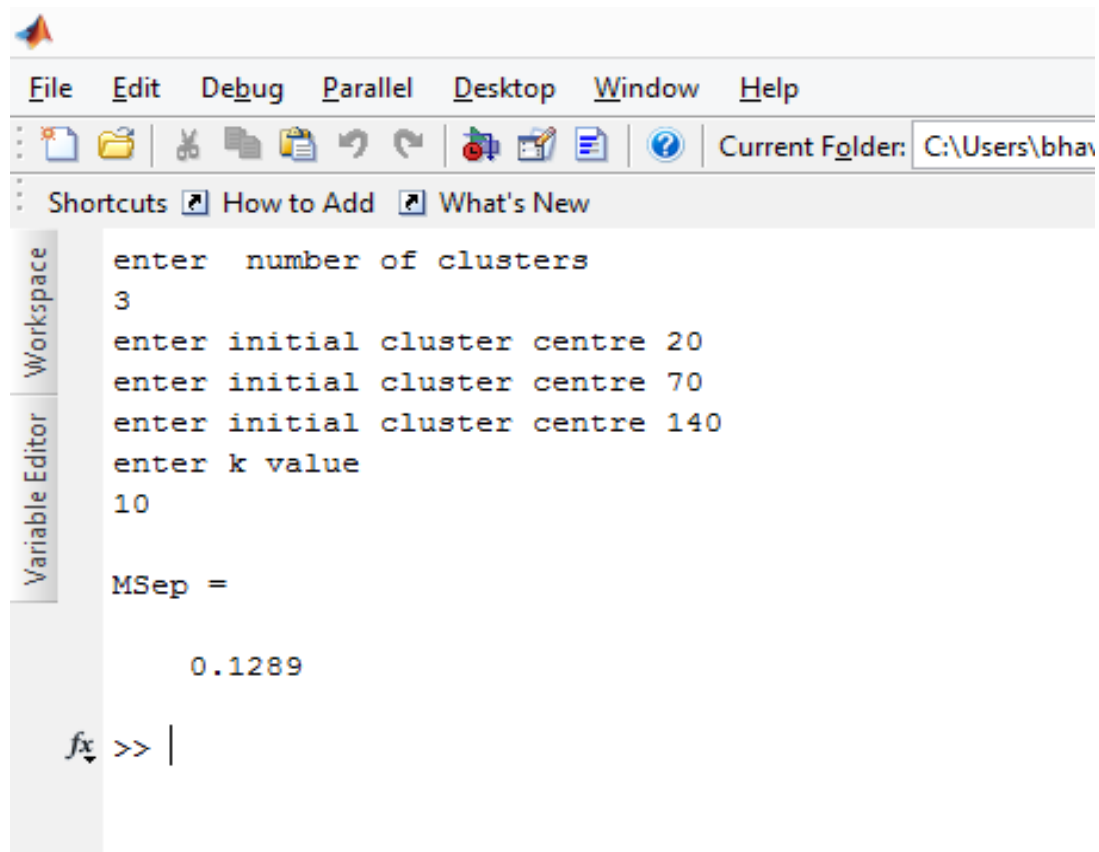
Workspace

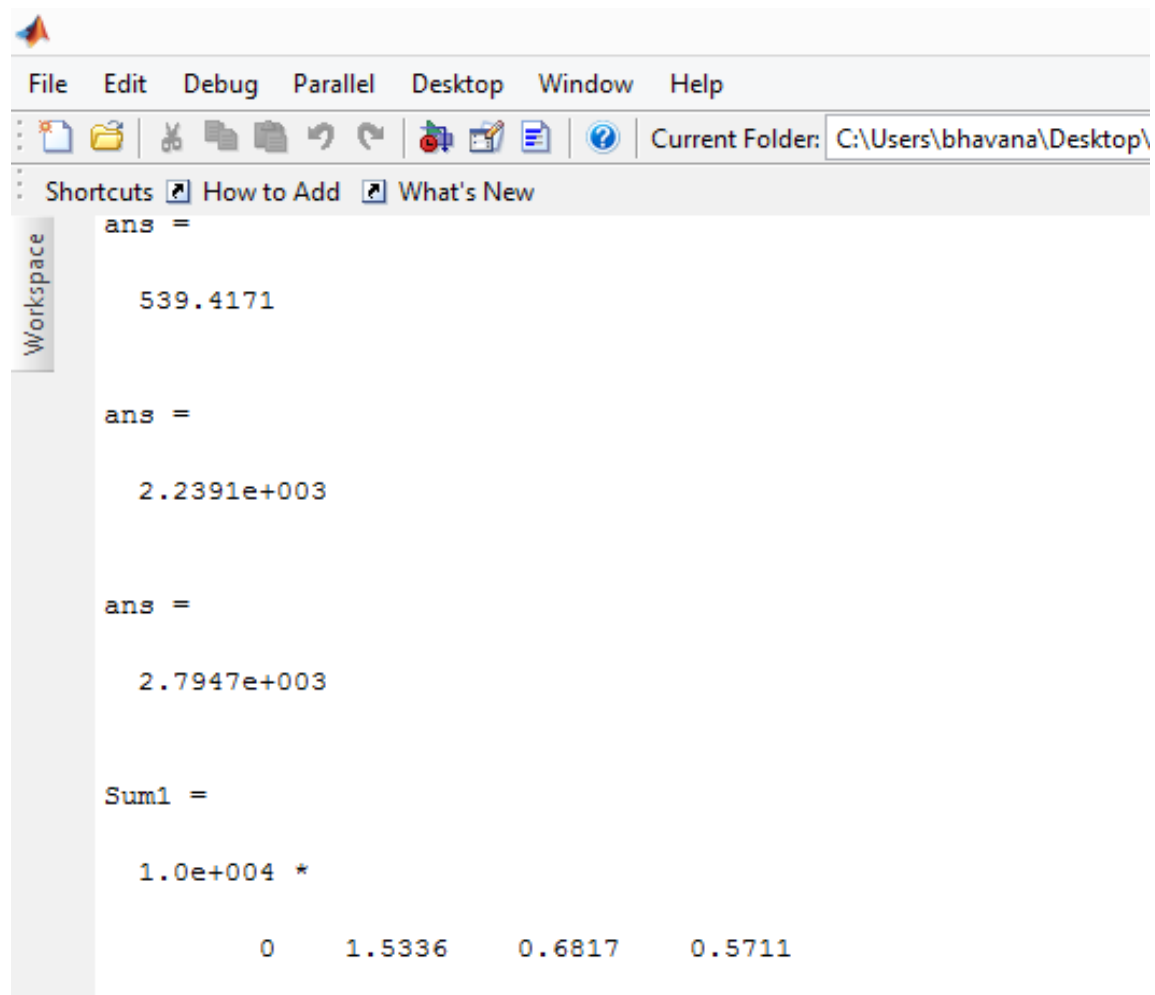
Variable Editor - Sep


Stack: Base No vali

Sep <1x3 double>










	1	2	3	4	5
1	0	0.1289	0.0982		
2					
3					
4					
5					
6					
7					
8					
9					









File Edit Debug Parallel Desktop Window Help

Current Folder: C:\Users\bhav

Shortcuts  How to Add  What's New

Workspace

```
ans =  
  
    539.4171  
  
ans =  
  
    2.0375e+003  
  
ans =  
  
    3.0430e+003  
  
Sum1 =  
  
    1.0e+004 *  
  
         0    1.5336    0.6817    0.5757  
  
enter k value  
10  
  
MSep =  
  
         0    0.0453    0.1289    0.2647
```

File Edit Debug Parallel Desktop Window Help

Current Folder: C:\Usr

Shortcuts How to Add What's New

Workspace

Variable Editor

```
ans =  
  
    539.4171  
  
ans =  
  
    2.0375e+003  
  
ans =  
  
    3.0430e+003  
  
Sum1 =  
  
    1.0e+004 *  
  
           0      1.5336      0.6817      0.5757  
  
enter k value  
10  
  
MSep =  
  
    0.1289      0.0453      0.1289      0.2647  
  
cvnn =  
  
           0      1.1711      0.9314      1.3754  
  
fx >>
```

Conclusion

- ▶ A new internal clustering validation measure, named CVNN is proved to be efficient in this project. Experimental results show that CVNN is capable to suggest the correct number of clusters on various synthetic and real-world data sets, including the data set with arbitrary cluster shapes.