Clusters using splitting and Merging

- Stability factor S_C of cluster C is a real number within [0, 1], used to measure quality of the cluster that represents proximity of the data objects within that cluster.
- Stability factor closer to 1 implies more stable the cluster is and of better quality.

Cluster using Splitting and Merging

- Here, intracluster distances are used for stability factors computation, based on which the clusters are splitted first.
- Later intercluster distances are calculated for merging of clusters.
- This iterative splitting and merging technique, finally provides stable clusters.

Splitting and Merging

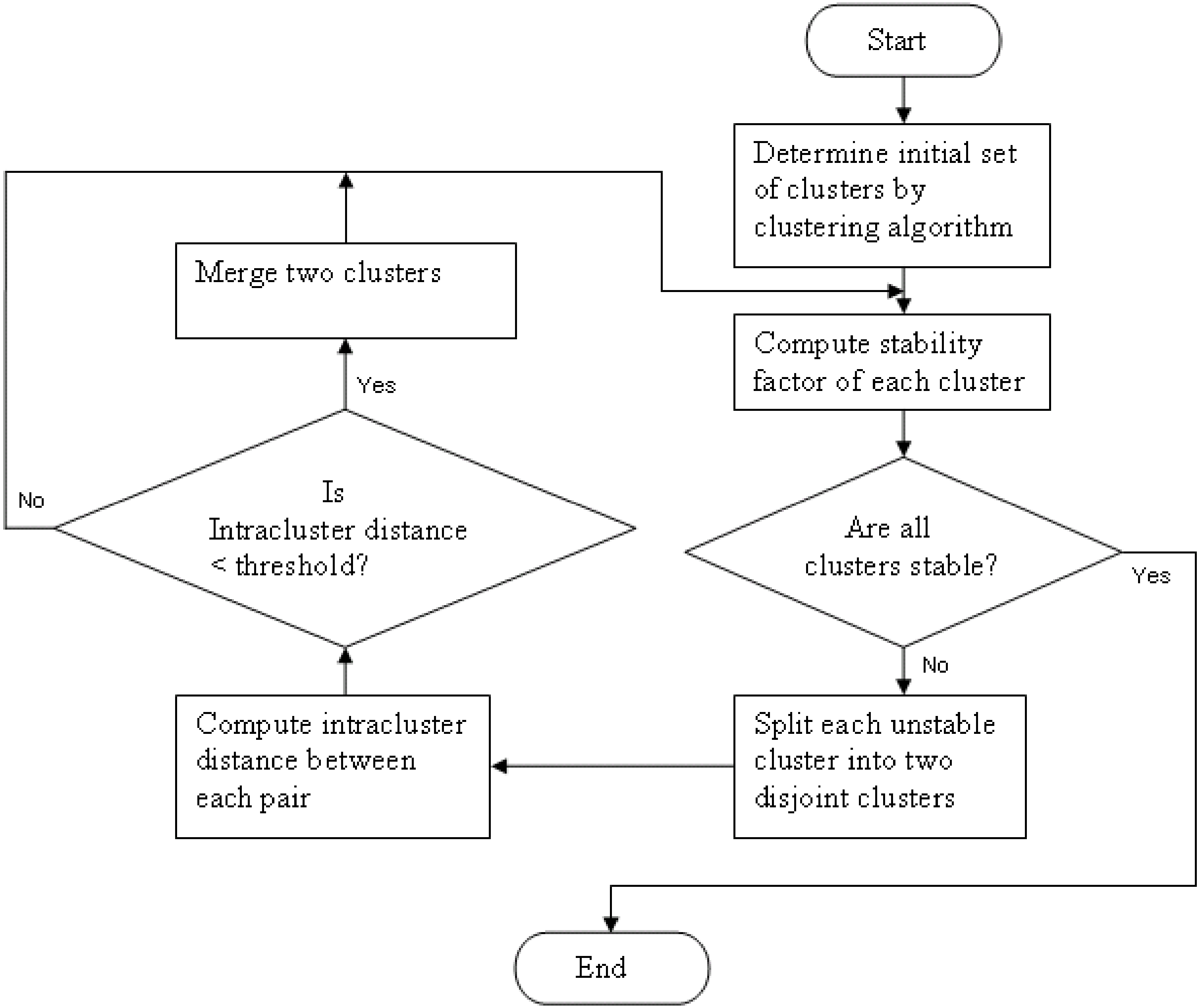
- Since, less the intracluster distance more close the objects are, so there is an inverse relationship between intracluster distance D_{CC} and stability factor S_C of cluster C.
- Satisfying this relationship, the inter stability factor of a cluster is computed using standard normal distribution function

$$S_C = \frac{1}{\sqrt{2\pi}} e^{-\mathrm{D}_{\mathrm{CC}}^{2}/2}$$

Splitting and Merging

Following basic steps are performed for cluster validation:

- For all clusters, intercluster and intracluster distances are computed.
- Stability factor of each cluster is computed using equation.
- If a cluster C is unstable (i.e., $S_C < \delta 1$, a threshold), then split it into two disjoint clusters.
- Merge two clusters S and D provided their intercluster distance $D_{SD} < \delta 2$, a threshold.
- This process is repeated till at least one splitting or merging of cluster take place. Thus the clusters are validated and the stable clusters are obtained.



Start
Determine initial set of clusters by clustering algorithm
Compute stability factor of each cluster
Are all clusters stable?
Yes
No
Split each unstable cluster into two disjoint clusters
Compute intracluster distance between each pair
Is Intracluster distance < threshold?
Yes
No
Merge two clusters
End

SPLIT - Algorithm

```
Procedure: SPLIT($C$, $l$)
Input: $l$, the number of data objects of cluster $C$.
Output: Two clusters $C1$ and $C2$.
Begin
    For $i$ = 1 to $l$ {
        For $j$ = 1 to $l$ {
            Compute $d_{ij}$ = the distance between $i$ and $j$
        }
    }
    For $i$ = 1 to $l$ {
        For $j$ = 1 to $l$ {
            Find $i$ and $j$ that maximize $d_{ij}$
        }
    }
    Form two clusters $C1$ and $C2$ with data objects $i$ and $j$ respectively
    For $k$ = 1 to $l$ {
            If ($dik \le djk$) then Insert data object $k$ into cluster $C1$
            Else Insert data object $k$ into cluster $C2$
    }
End.
```

Results

- The Electronic shop dataset is divided into nine disjoint clusters by the *SAM*-algorithm