**Estimation of number of cluster**

Overview of cluster analysis

* A Survey of Clustering Algorithms for Big Data: Taxonomy and Empirical Analysis

Evaluation of clustering model (validity evaluation)

* External indices – comparing with external information true cluster labels
  + Classification Accuracy (CA)
  + Adjusted Rand Index (ARI)
  + Normalized Mutual Information (NMI).
  + F-measure
  + Purity
  + Entropy
* Internal indices – measure the compactness and separation of the clusters
  + Compatness (CP)
  + Dunn Validity Index (DVI)
  + Distance-based
    - Caliński T, Harabasz J (1974) A dendrite method for cluster analysis. Commun Stat Theory Methods 3(1):1–27
    - 9. Krzanowski WJ, Lai Y (1988) A criterion for determining the number of groups in a data set using sum-of-squares clustering. Biometrics 44(1):23–34
    - Rousseeuw PJ (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math 20:53–65
    - Tibshirani R, Walther G, Hastie T (2001) Estimating the number of clusters in a data set via the gap statistic. J R Stat Soc Ser B (Stat Methodol) 63(2):411–423
  + Depth-based
    - Estimating the Optimal Number of Clusters k in a Dataset Using Data Depth

Approach of determination of cluster number

1. Traditional approach

AIMS: summarize and outline methods for selecting number of clusters

**Paper Title:**

Recovering the number of clusters in data sets with noise features using feature rescaling factors, introducing

1. Citations
   1. Recovering the number of clusters in data sets with noise features using feature rescaling factors
   2. **[15] M. Halkidi, Y. Batistakis, M. Vazirgiannis, On clustering validation techniques, J. Intell. Inf. Syst. 17 (2-3) (2001) 107–145.**
   3. **M.M.-T. Chiang, B. Mirkin**
   4. **Intelligent choice of the number of clusters in k-means clustering: an experimental study with different cluster spreads**

**CVIs**

1. **Silhouette index –** normalized difference of dissimilarity over within and between clusters
2. **Dunns index** – ratio betweensmallest distancae between clusters and the maximum cluster diameter
3. **Clinski-Harabasz index** – a ratio of a between-cluster-means and a within-cluster sum of squares statistic
4. **Hartigan index -**

K-means variation (2000~)

1. X-means
2. Ik-means (Intelligent k-means)
3. Fuzzy c-means

Kmeans estimation of number of cluster

Year 2010~

1. MMT Chiang 2010, Intelligent Choice of the Number of Clusters in K-Means Clustering: An Experimental Study with Different Cluster Spreads [231 citation]

Year 2015~

1. Fuzzy clustering
2. Deep-learning based dimension reduction prior to clustering
   1. B Yang et al, 2017, Towards k-means-friendly spaces: Simultaneous deep learning and clustering [202]
   2. V.Patel et al 2013, Latent Space Sparse Subspace Clustering [130]
3. **Fuzzy-based**
4. Spectral Clustering algorithm – グラフ理論  
   AngelMur, 2016
5. MA Masud et all 2018, I-nice: A new approach for identifying the number of clusters and initial cluster centres [43 citation]
6. **Density-based**

-