- Sim function => Categorical value নিয়ে কাজ করে।
- Dis function => Numerical value নিয়ে কাজ করে।
- Nearest Neighbor clustering local optima তে stuck হয় না।

Define Ensemble Clustering

Ans:

It's a Process of integrating multiple clustering algorithms to form a single strong clustering approach that usually Provides better clustering results.

Write the advantages of Ensemble Clustering

Ans:

The advantages of Ensemble clustering are

a. Performs more effectively in high dimensional Complex data.

b. Good alternative when facing cluster analysis Problems.

Write down the strategies of Ensemble Clustering

Ans:

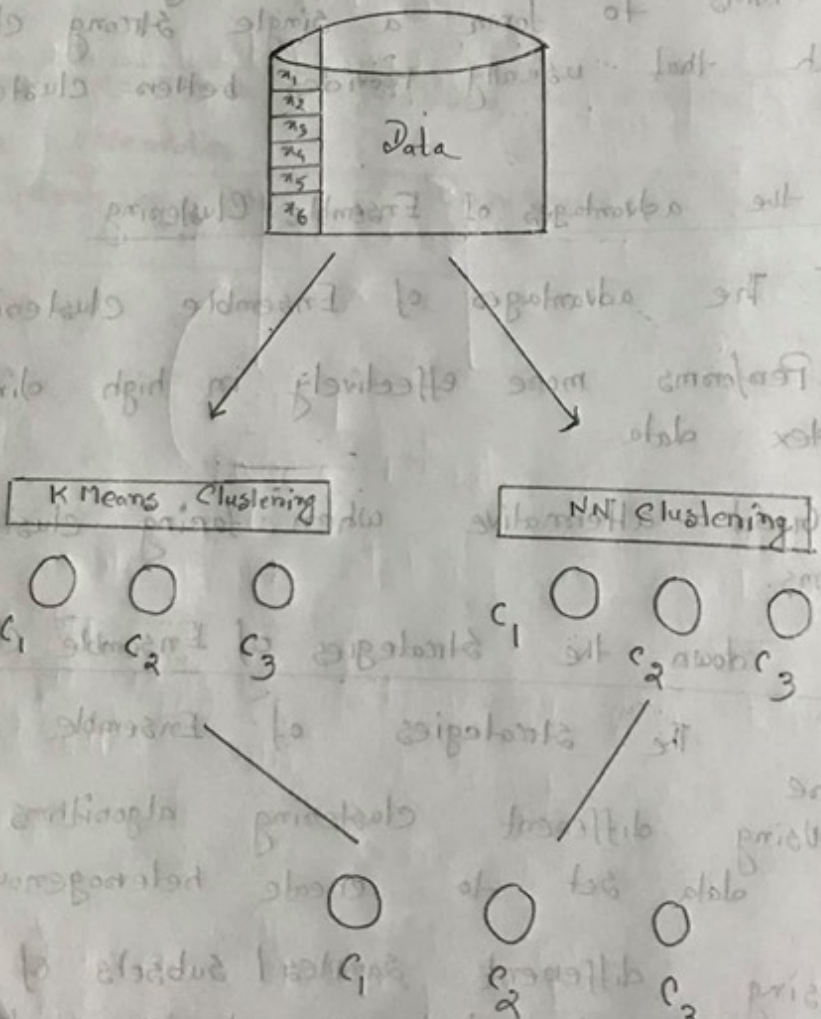
The strategies of Ensemble clustering are

a. Using different clustering algorithms on the same data set to create heterogenous clusters.

b. Using different samples / subsets of the data with different clustering algorithms to cluster them to Produce Component clusters.

c. Running the same clustering algorithms many times on same data set with different Parameters or initialisations to create homogenous clusters.

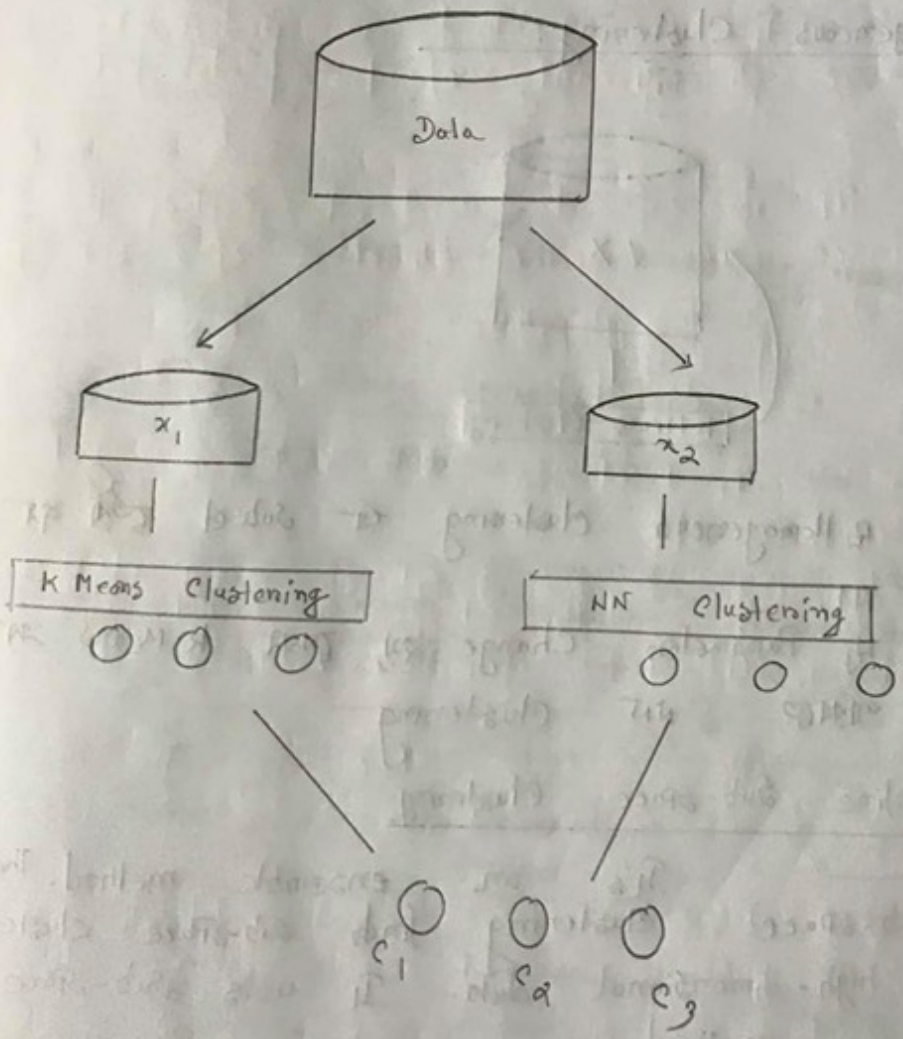
Helenogenous Clustering:



Heterogeneous Clustering হচ্ছে →

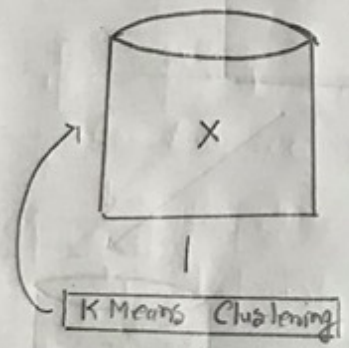
- Model একই data এর উপর train করা হয়,
- Serial দেখে বুঝা যায় একটা instance কোন Model এর কোন cluster থেকে আসছে,

Component Clustering:



- Component Clustering এর বেলা বলায়, instances গুলোর কোন serial নাম্বার থাকে না।
- Component clustering এর ক্ষেত্রে বড় dataset কে sub-dataset এ ভাগ করা হয়, Equally sub-data
- Component clustering এ, Cluster এর Property similar কিনা দেখা হয়,

Homogeneous Clustering:



- Homogenous clustering এ subset তৈরা হয় না,
- Parameter Change হয়, যেমন k means এর পরিবর্তে NN Clustering

- Define Sub-space Clustering.

Ans:

Sub-space clustering is an ensemble method. The finds sub-space clusters in high-dimensional data. It uses sub-space search Method.

Sub-Space ⟶ Sub Space Search Method

✠ $X = \{A_1, A_2, A_3, \dots A_n\}$

↓ full - space

A_1, A_2 → এগুলো প্রতি Dimension / Feature.

| A_1 | A_2 | A_3 | A_4 | A_5 | ⇒ 5 Column / Input Feature

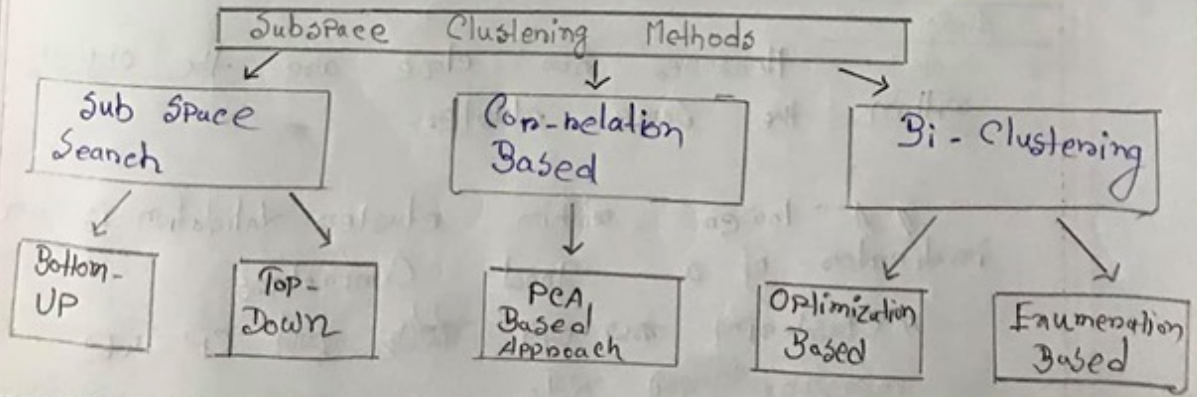
✠ Sub-Space ⇒ এটা Sequentially অথবা Randomly যেকোন হয়।

✠ Define Sub Space Search Clustering

Ans:

Search various subspaces to find clusters using

i. Bottom up approaches : low to high dimension.
ii. Top down approaches: full space to smaller subspaces recursively.



Define Co-relation Based Clustering:

Ans:

It uses space transformation methods to derive a set of new, uncorrelated dimensions & then mine clusters in new space on its subspaces. It uses PCA.

What is PCA (Principal Component Analysis)

Ans:

It's a method to reduce the ~~dimensio~~ dimensionality of large data sets.

It is manily used to explain the variance - co-variance structure of a set of variables through linear combinations.

What is Bi-Clustering

Ans:

It clusters both instances & features simultaneously.

// Bi-Clustering → Row ও Column একসাথে clustering করা হয়। একে delta clustering ও বলে,

What is (CP) Cluster Compactness / Cluster Cohesion

Ans:

Measures how close are the objects within the same cluster.

// A lower within cluster validation is an indicator of a good Compactness.

// Clustering এর ভাল কতটুকু হয়েছে তা CP দিয়ে measure করা হয়,

Define Ensemble classifier.

Ans: It's a process of combining different classification techniques to build a powerful composite model from the data.

- Ensemble classification হল supervised learning.
- Ensemble খুবই class imbalance data এর জন্য ভালো কাজ করে।

Draw Ensemble classifier model.

Ans.

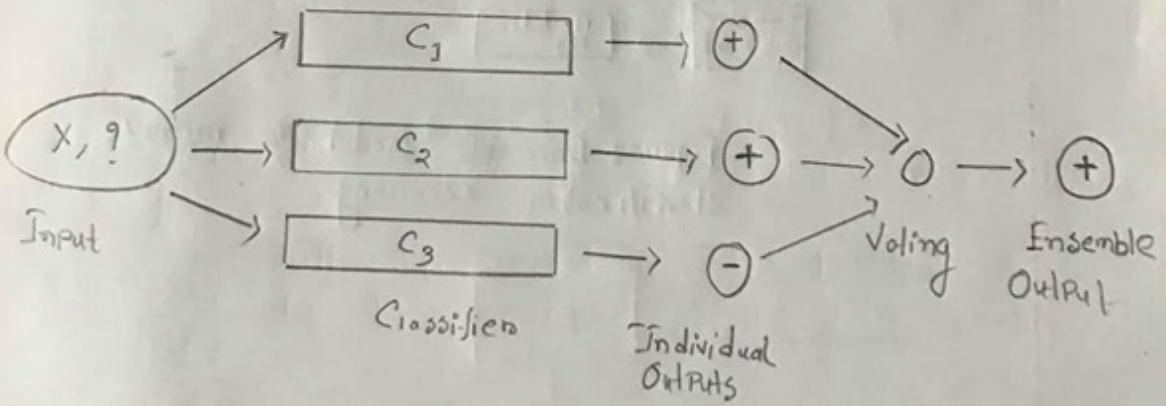


Figure: Ensemble classifier

What is the advantage of training data in Ensemble Classifier

Ans

From training data, derive several sub-data sets from it. Then learn a classifier/model from each. Later combine them to produce an ensemble model.

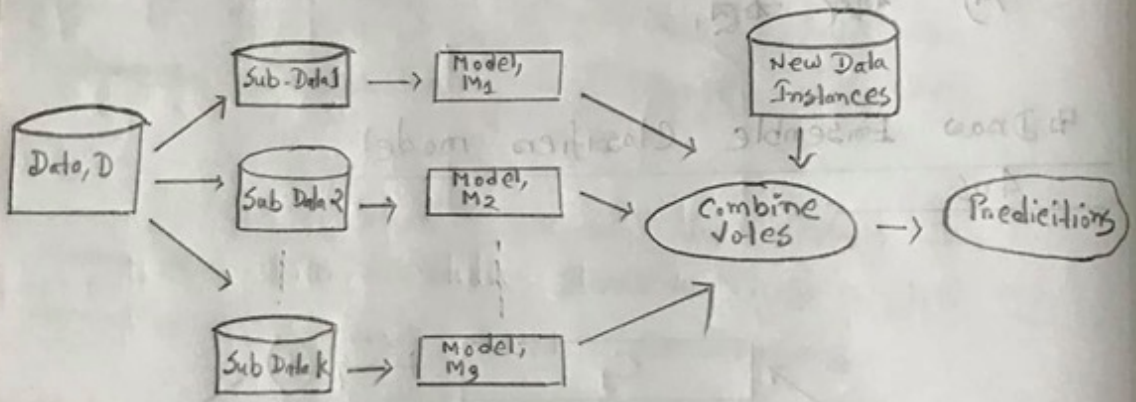


Figure: Ensemble model to improve classification accuracy.

Describe Random forest with diagram

Ans:

Random forest is an ensemble learning method for classification & regression that able to classify large amounts of data with accuracy.

It constructs a number of decision trees based on randomly selected sub-sets of features at training time. It outputs class that is the ensemble of trees vote for the most popular class.

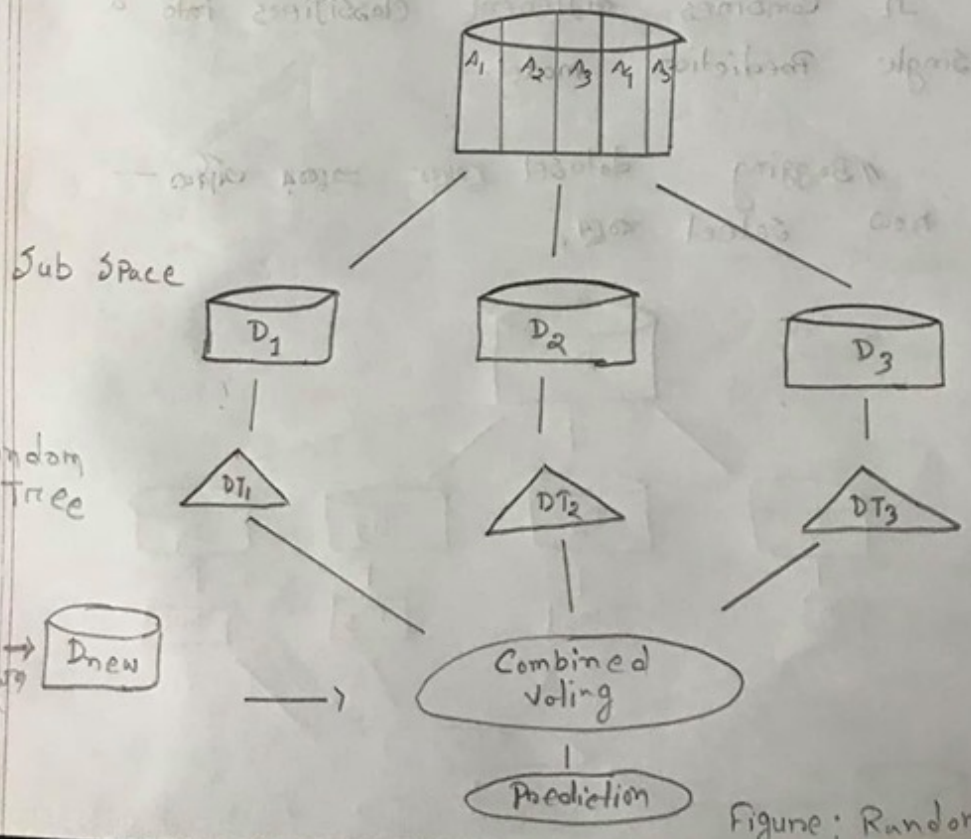


Figure: Random Forest

Page

- Random forest এক্ষেত্রে অধিক decision tree নিয়ে কাজ করে।
- প্রতিবার decision tree random column এর উপর base করে তৈরী হয়।

- Regression Problem => Class label Numeric
- Classification Problem => Class label Categorical.

Define Bagging

Ans: Bagging means Bootstrap Aggregation. It combines different classifiers into a single Prediction model.

// Bagging dataset থেকে একের অধিক — row select করে।

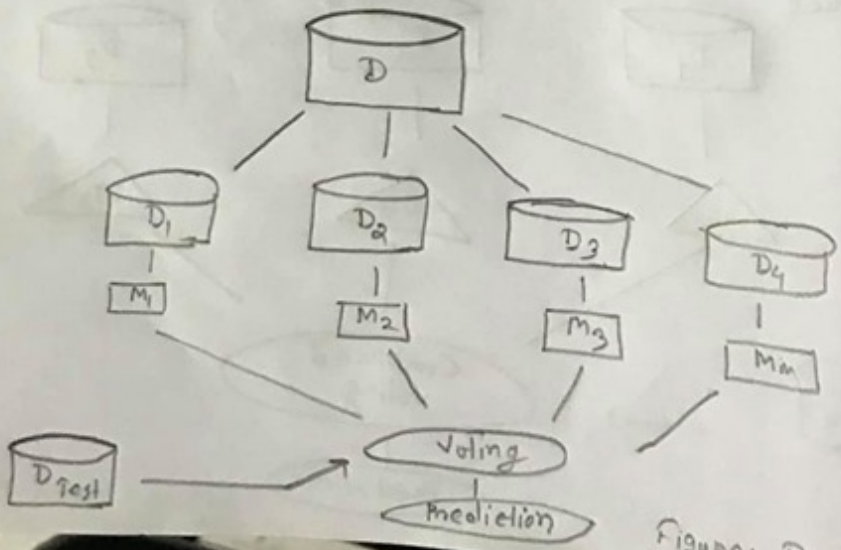


Figure: Bagging

Write Bagging Algorithm

Ans

Input: Training data D, Number of Iterations K, A learning scheme

Output: Ensemble Model, M^*

Method:

1. for $i = 1$ to k do
2. Create bootstrap sample D_i, by sampling D with replacement;
3. Use D_i & learning scheme to derive a model, M_i;
4. end for.

To Use M^* to classify a new instance, X_{new}:
Each $M_i \in M^*$ Classify X_{New} & return the majority vote

$M^* \Rightarrow M_1, M_2, M_3$ এগুলো OneR / SVM / KNN দিয়ে বানানো model. একই অথবা ভিন্ন হতে পারে।