Q What is Ensemble Trees.

Ans:

It combines a number of decision trees in order to reduce the risk of overfitting

Q Write Ensemble Tree Algorithm

Ans:

Input: Training Data, D, $C_{4.5}$ learning Algorithm

Output: A set of trees, $DT^*$

Method:

1. $DT^* = \phi$
2. Create sub-datasets, $D_1, \ldots, D_i, \ldots D_k$ from the training data, D;
3. for i=1 to k do
4. group features in $D_i$ into m groups;
5. for j=1 to m do
6. build a $DT_j$ with $j_{th}$ feature group;
7. complete error($DT_j$) on $D_i$ ;
8. if error($DT_j$) ≤ threshold_value then
9. $DT^* = DT^* \cup DT_j$;
10. end if
11. end for
12. end for

## Define Class Imbalanced Problem

Ans:

It is the Problem in machine learning where the total number of a class of data (Positive) is far less than total numbers of another class of data (negative).

## Write the Causes of overfitting & Overlapping of Classes

Ans: The causes of overfitting & overlapping are

a. Noise
b. High dimension
c. Small Sample size.

## Which methods are used for class imbalancing

Ans:

Three methods are used in class imbalancing. They are

a. Sampling Method : i. Under Sampling
ii. Over Sampling

b. Cost - Sensitive Learning Method

c. Ensemble Methods : i. Bagging
ii. Boosting

## Describe Under - Sampling & Over Sampling

## Under Sampling:

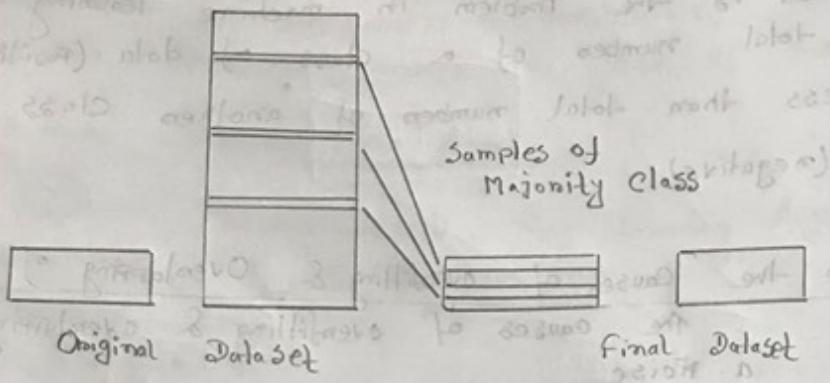


Figure: Under Sampling

By Undersampling

a. Runtime can be improved by decreasing the amount of training dataset.

b. Helps to solve the memory Problem.

It causes information loss.

## Over Sampling:

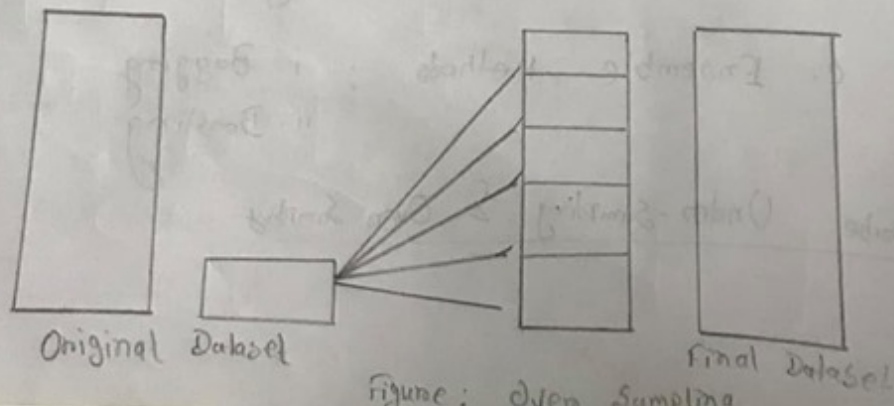


Figure: Over Sampling

By Over Sampling

a. No loss of information
b. Mitigate overfitting caused by oversampling

What is Cost Sensitive Learning. Write the advantage

Ans:

It takes the mis-classification costs into consideration by minimising the total cost. The goal of this technique is to ~~Pursue~~ gain a high accuracy of classifying examples into a set of known classes.

Advantage

a. Avoids Pre-selection of Parameters.
b. Auto adjust decision hyperplane.

## Explain Data Balancing Method

Ans:

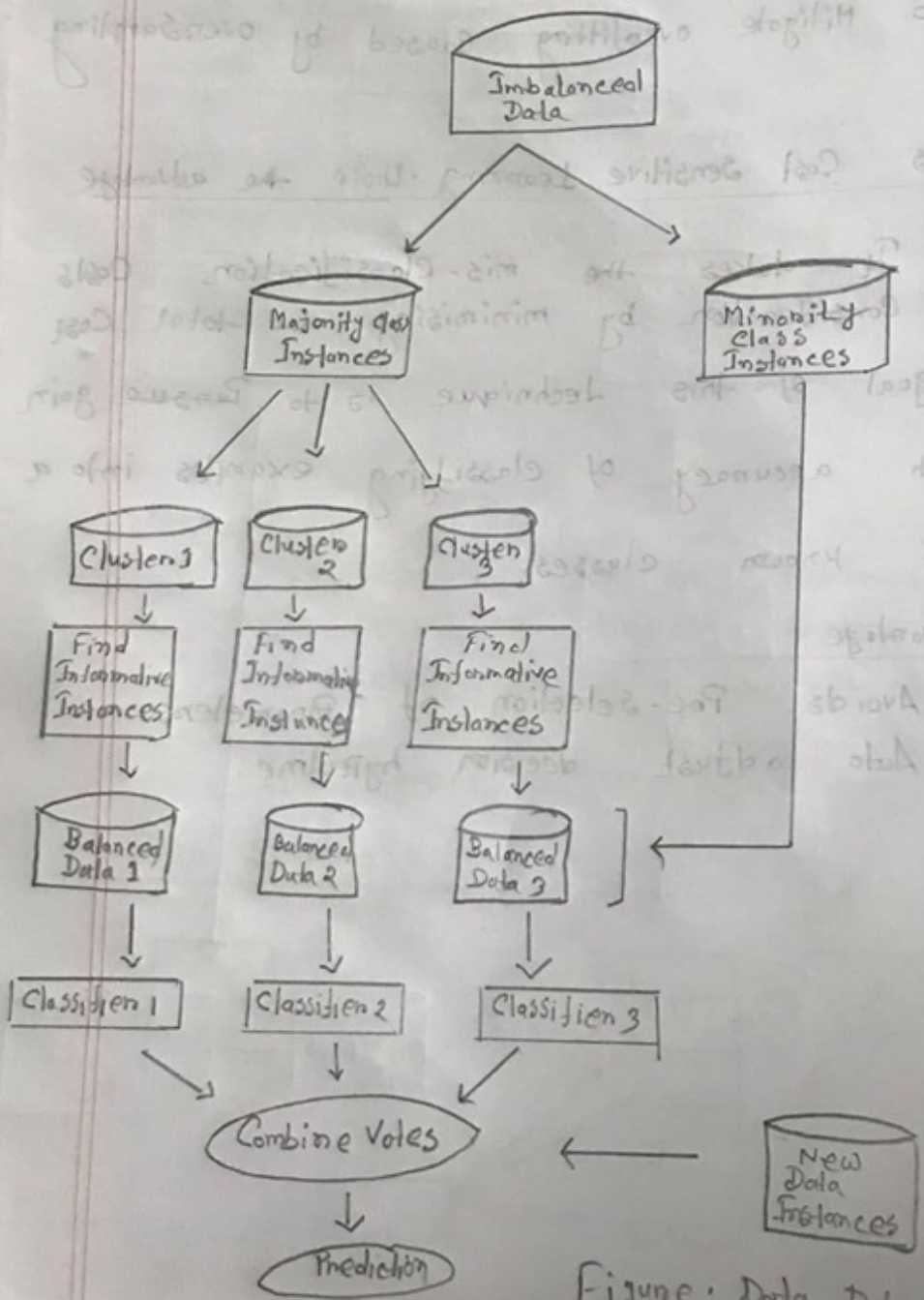


Figure: Data Balancing

For Data Balancing

a. Majority class instances are clustered into several clusters.

b. Next comes finding most informative instances in each cluster.

c. Informative instances are close to the center or border of the cluster.

d. With these most informative instances several datasets are created.

e. Every dataset should have almost equal number of minority - majority classes instances.

f. Multiple classifiers are trained using these datasets

Page

## Explain Active learning:

Ans.

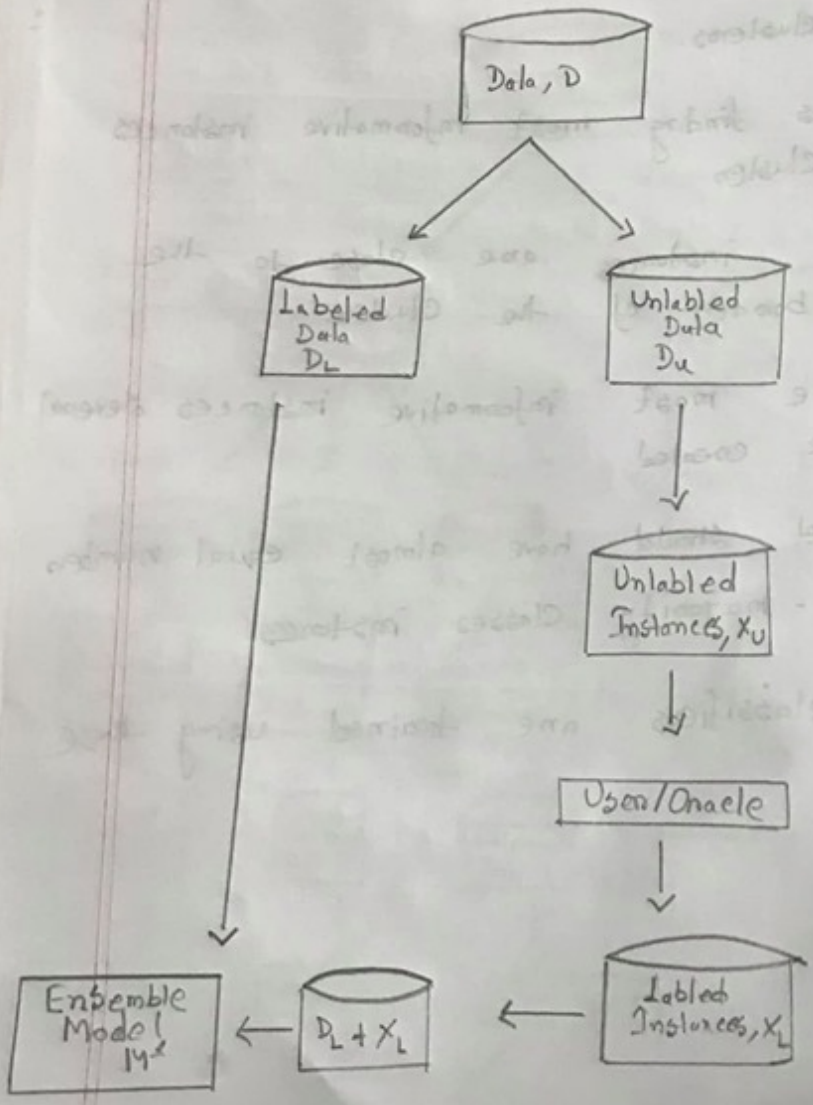


figure: Active learning

An active learning can ask a user to label an instance, which can be from a set of unlabelled instances.

Naive Bayes classifier & clustering are used to find the most informative instances for labeling as part of active learning. It follows two key rules

a. Instances close to centers of clusters & border of clusters.

b. Posterior Probabilities of instances are equal / very close.

Q. Write the steps of Data Pre-Processing

Ans:

The steps of Data PreProcessing are;

a. Data Cleaning: Dealing with missing values.

b. Data Integration: Merge multiple sources data into Coherent data store.

c. Data Transformation: It includes four sub steps

i. Normalisation
ii. Aggregation
iii. Generalisation
iv. Feature Construction.

d. Data Reduction: Eliminating redundant features.

e. Data Discretisation: Reduction of a number of values of a continuous feature.

☆ What is feature selection. Write three reasons of feature selection

Ans:

Feature selection means the Process of selecting a subset of relevant features

The three reasons of feature selection are

a. Simplification of models
b. Shorten training times
c. Reducing overfitting.

☆ What is Entropy

Ans:

The measure of randomness in the information being Processed.

The End

30/08/2019