Novel application of Random Forest method in CERES scene type classification

Bijoy V. Thampi[1]
Constantine Lukashin[2]
Takmeng Wong[2]

[1]Science System Applications Inc., Hampton, VA
[2]NASA Langley Research Center, VA

CERES Science Team Meeting
Scripps Institution of Oceanography, San Diego, October 29-31

Objective of the study

The motivation for this study is to develop a machine learning method for an improved estimate of ERBE like fluxes from instruments on spacecraft that have no imager data.

The methodology can be used to infer

- TOA fluxes when there is insufficient imager coverage
- TOA fluxes when there is an imager failure
- Classify scene type using CERES radiance and available ancillary data.

Machine learning

Machine learning focuses on model prediction, based on known properties learned from the training data.

Ensemble learning is a machine learning paradigm where multiple models (learners) are trained to solve the same problem. By using multiple learners, generalization ability of an ensemble can be much better than single learner.

Main advantages of Ensemble learning methods are are:

Reduced variance: *results are less dependent on peculiarities of a single learner and training set..*

Reduced bias : *combination of multiple classifiers may produce more reliable classification than single classifier.*

Eg.: Boosting, Bagging, Random forest, stacking...

Random Forests

Random forest, first proposed by Tin Kam Ho of Bell Labs in 1995, is an ensemble learning method for **classification and regression**

The algorithm for inducing a random forest was developed by Leo Breiman and Adele Cutler (2001)

Main idea : build a larger number of decision trees(base learners)

Motivation : reduce error correlation between classifiers

Key : using a random selection of features to split on at each node

Advantages:

RF is easy to build and faster to predict!

Resistance to over training and over-fitting of data

Ability to handle data without preprocessing or rescaling.

Resistant to outliers and can handle missing values.

Random Forests

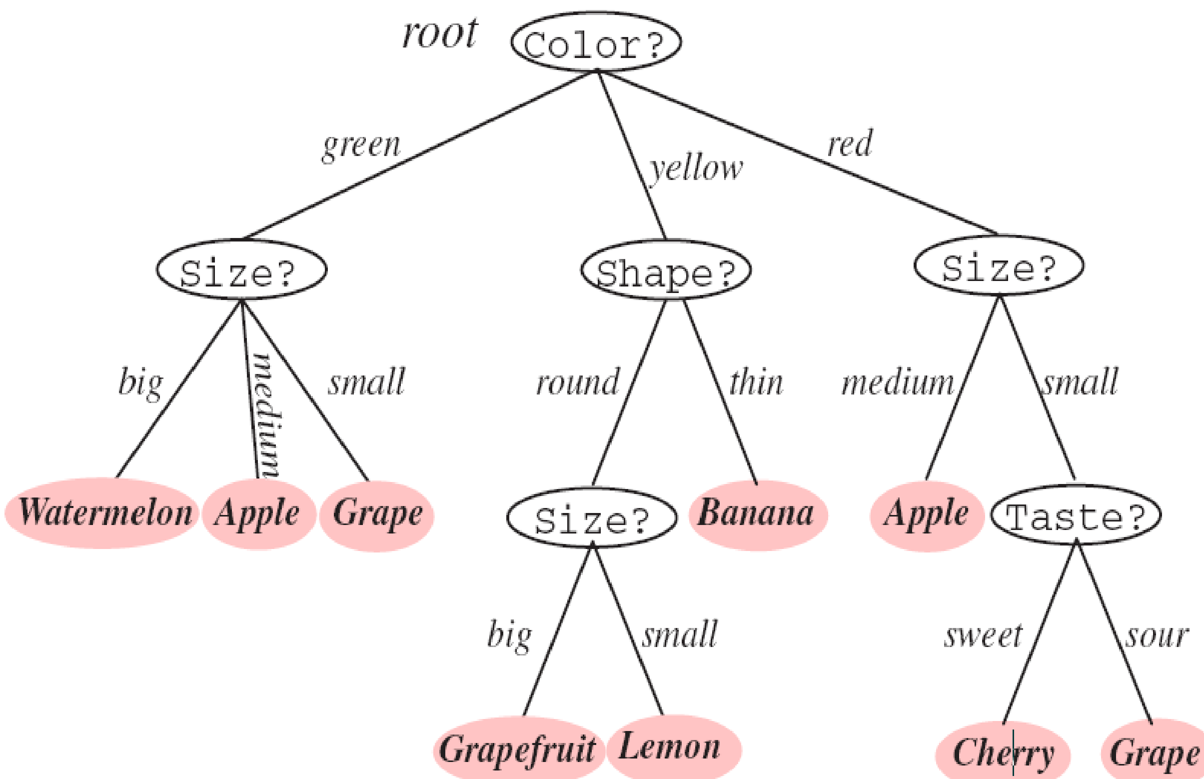
- Use **decision tree classifiers** as the base learner

  A flow-chart-like tree structure

  Internal node denotes a test on an attribute

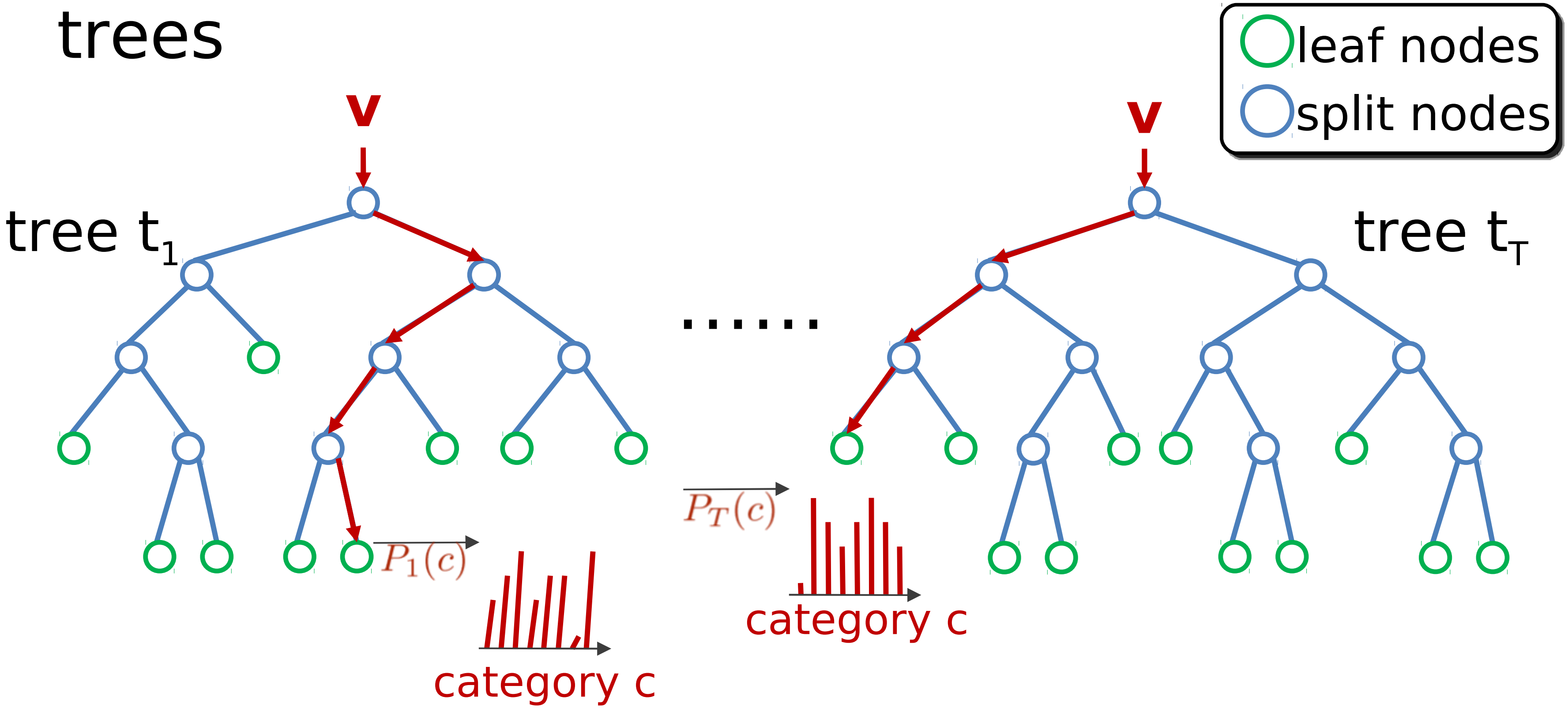
  Branch represents an outcome of the test

  Leaf nodes represent class labels or class distribution



A Forest of Trees

- Forest is an ensemble of several decision trees



$$P(c|\mathbf{v}) = \sum_{t=1}^{T} P_t(c|\mathbf{v})$$

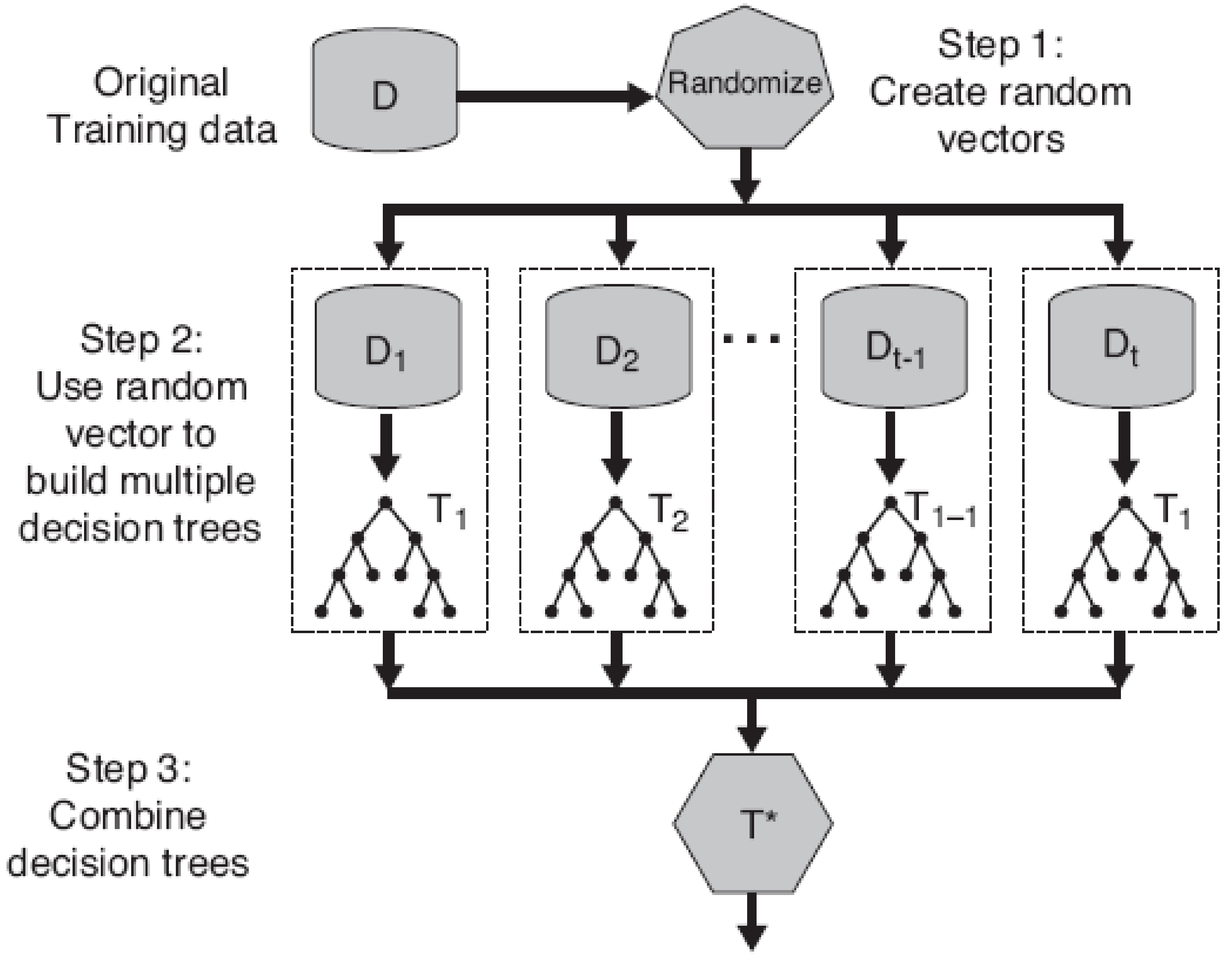
P(c|v) - final classification of forest
P_t(c|v) - classification at each tree
T - Number of trees built

Random Forest Algorithm

(Brieman and Cutler, 2003)

- Introduce two sources of randomness: “**Bagging**” and “**Random input vectors**”.
  - **Bagging**- creating ensembles by “bootstrap aggregation”- repeated random sub-sampling of the training data.
  - **Bootstrap sample** - will on average contain **63.2%** of the data while the rest are replicates.
- Using bootstrap sample, a decision tree is grown to its greatest depth minimizing the loss function.
- At each node, best split of decision tree is chosen from **random sample of input variables** instead of all variables.
- For each tree, using the leftover (36.8%) data, calculate the misclassification rate = **out of bag (OOB)** error rate.
- Aggregate error from all trees to determine **overall OOB error rate for the classification**

Random Forests –Flow diagram



Classification of CERES Scene type

Objective: to classify scene types using CERES radiance and ancillary data.

Our primary goal was to test the efficiency of RF in classifying the CERES radiances as **clear and cloudy**.

Initially, the training dataset is labelled (radiances are classified as clear and cloudy) while the test dataset is unlabelled. Using the trained forest, classes of the test dataset are predicted.

The main steps involved in the RF scene classification are:

- Definition of the training and test datasets
- Supervised training of random forest on the training sets.
- Classification of the test data using the saved forest.
- Error determination

RF – Input variables

Input variables are selected for the scene classification are:

CERES

- solar zenith angle & viewing zenith angle
- relative azimuth angle
- CERES LW and SW broadband radiances
- IGBP Surface type

Ancillary (Reanalysis)

- LW surface emissivity
- Broadband surface albedo
- Surface skin temperature
- Column averaged relative humidity
- Precipitable water

Training & Test data

Class No	Surface type	Scene type	Number of samples	
			Training	test
1	Water	Clear	**11230**	**11230**
2	Water	Cloudy	**11430**	**11430**
3	Bright desert	Clear	**5545**	**5476**
4	Bright desert	Cloudy	**7654**	**7590**
5	Dark desert	Clear	**10025**	**10200**
6	Dark desert	Cloudy	**10789**	**10750**
7	Snow	Clear	**6810**	**6940**
8	Snow	Cloudy	**9117**	**9058**

Source: **CERES Terra SSF**

Training data: **July 2003**

Test data : **July 2004**

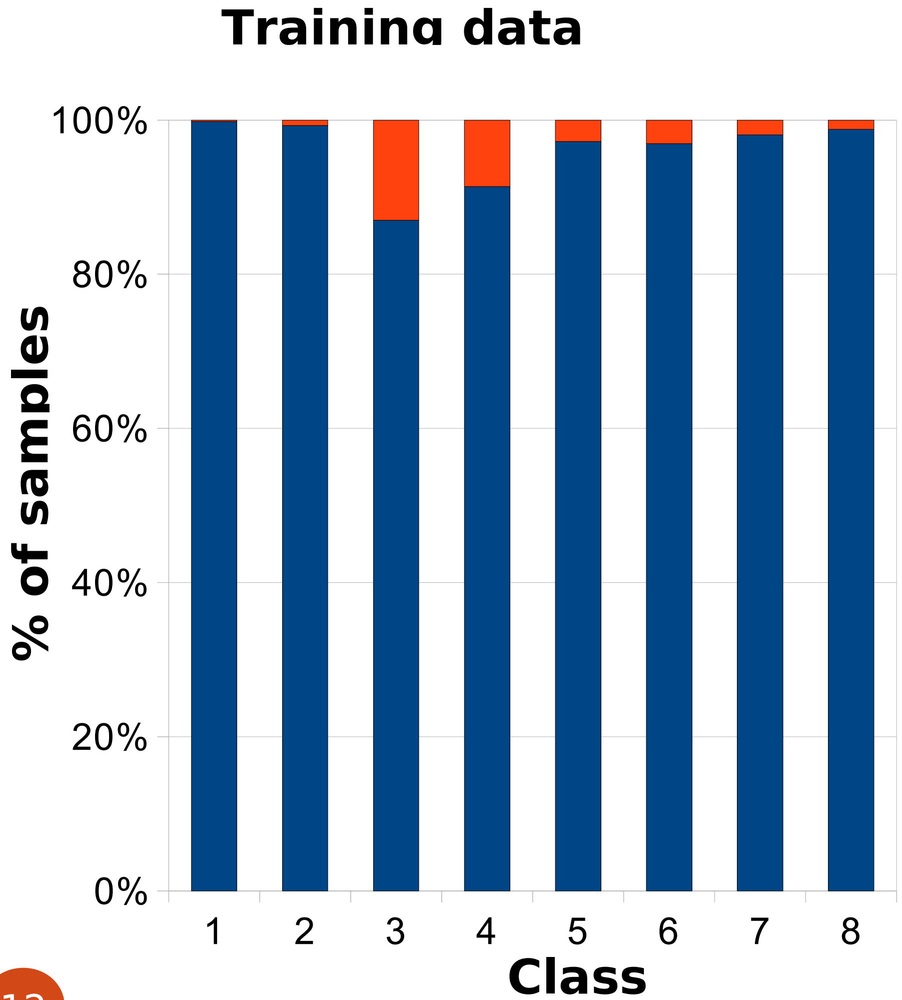
CERES SSF dataset contains millions of CERES footprints.

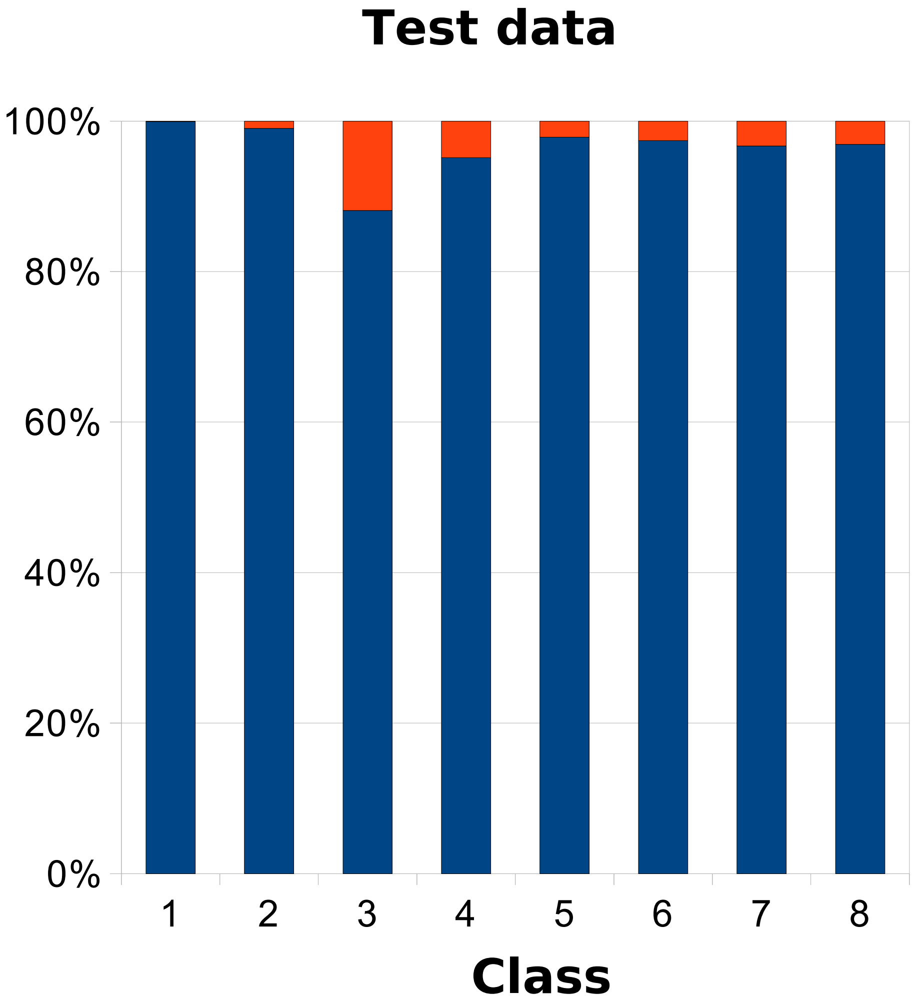
Need to cereate compact training sets.

This is achieved by stratifying the data in the variable of interest (SZA, VZA, RZA)

RF Scene classification – Results

Bluish shade represent correct classification of clear /cloudy
Orange shades represent Incorrectly classified data samples





RF Scene classification - Error Analysis

Number of input variables : 10
Number of trees built : 500

RF Classification Error (%) associated with each class

CLASS	1	2	3	4	5	6	7	8
Training	0.18	0.7	13	8.6	2.8	3.1	1.9	1.2
Test	0.1	1.0	11.9	4.9	2.2	2.6	3.3	3.1

Final error rate (%)

Training set : **3.2**

Test set : **3.0**

RF Scene Classification- ERBE like

In this analysis, Scene classification is performed using the random forest with only **ERBE like** variables as input.

Number of input variables : 5
Number of tress built : 500

Classification Error (%) associated with each class

CLASS	1	2	3	4	5	6	7	8
Training	5.5	10	29.4	48.1	9.1	25.3	10.1	11.9
Test	3.3	5.9	21.8	49.8	22.9	26.4	12.8	21.1

Final error rate (%) Training set : **17.2 (3.2)**
Test set : **19.2 (3.0)**

Conclusions

- Random forest is one of the most advance ensemble learning algorithms available and is a highly flexible classifier.
- It runs efficiently on large databases.
- RF classification of CERES Scene types (Clear and cloudy) shows very good classification of clear and cloudy radiances with avg. error <5 % over most surface types.
- Scene classification error shows considerable increase >10% for most scene types when ancillary variables are removed (ERBE like approach).
- **Future Plans:**

  Expand the database including multiyear

  Expand the scene classes- cloudy water, cloudy ice,…

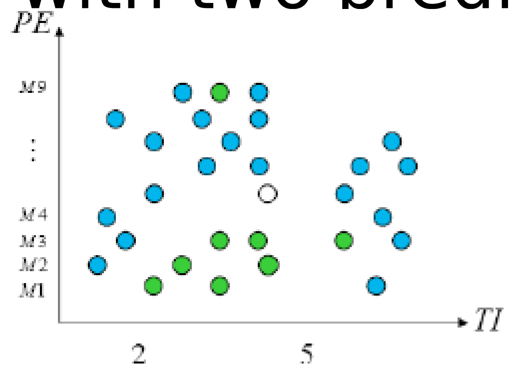
  Include more non imager variables for better classification


THANK YOU VERY MUCH
FOR LISTENING!!!
ANY QUESTION?

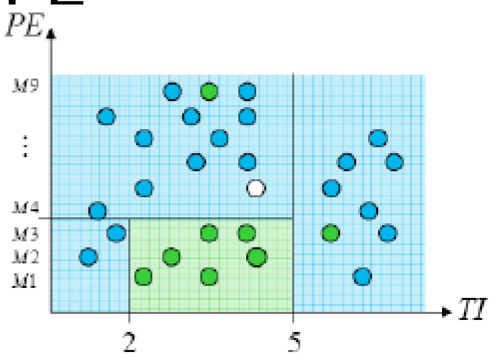
Decision tress involve greedy, recursive partitioning.

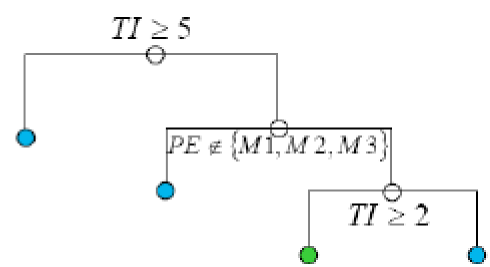
- Simple dataset with two predictors

TI	PE	Response
1.0	M2	good
2.0	M1	bad
...	...	...
4.5	M5	?



- Greedy, recursive partitioning along TI and PE





Scene classification error

	Training data (March 2003)	Test data (March 2004)
samples size	63700	60120
classes	8	8
Variables	10	10
Variable split at node	3	3
Decision tress grown	500	500

Classification Error (%) associated with each class

Class	1	2	3	4	5	6	7	8
Training set	0.4	0.3	5.5	2	7.9	6.9	1.5	1.5
Test set	0.1	2.8	2.6	3.2	3.7	11.5	1.6	6.2

Final error rate (%) Training set : **2.9**

Test set : **3.8**

Classification error - ERBE like

Classification Error (%) associated with each class

Class	1	2	3	4	5	6	7	8
Training set	6.7	7.6	11.2	35.8	21.2	51.2	22.9	6.05
Test set	1.2	7.0	22.6	54.4	12.8	54.9	43.2	22.11

Final error rate (%) Training set : **19.5 (2.9)**
Test set : **25.0 (3.8)**