Random Forest

Data Base and Data Mining Group of Politecnico di Torino

Elena Baralis

Politecnico di Torino

Random Forest

- Ensemble learning technique
  - multiple base models are combined
    - to improve accuracy and stability
    - to avoid overfitting

- Random forest = set of decision trees
  - a number of decision trees are built at training time
  - the class is assigned by majority voting

Bibliography: Hastie, Tibshirani, Friedman, The Elements of Statistical Learning, Springer, 2009

Random Forest

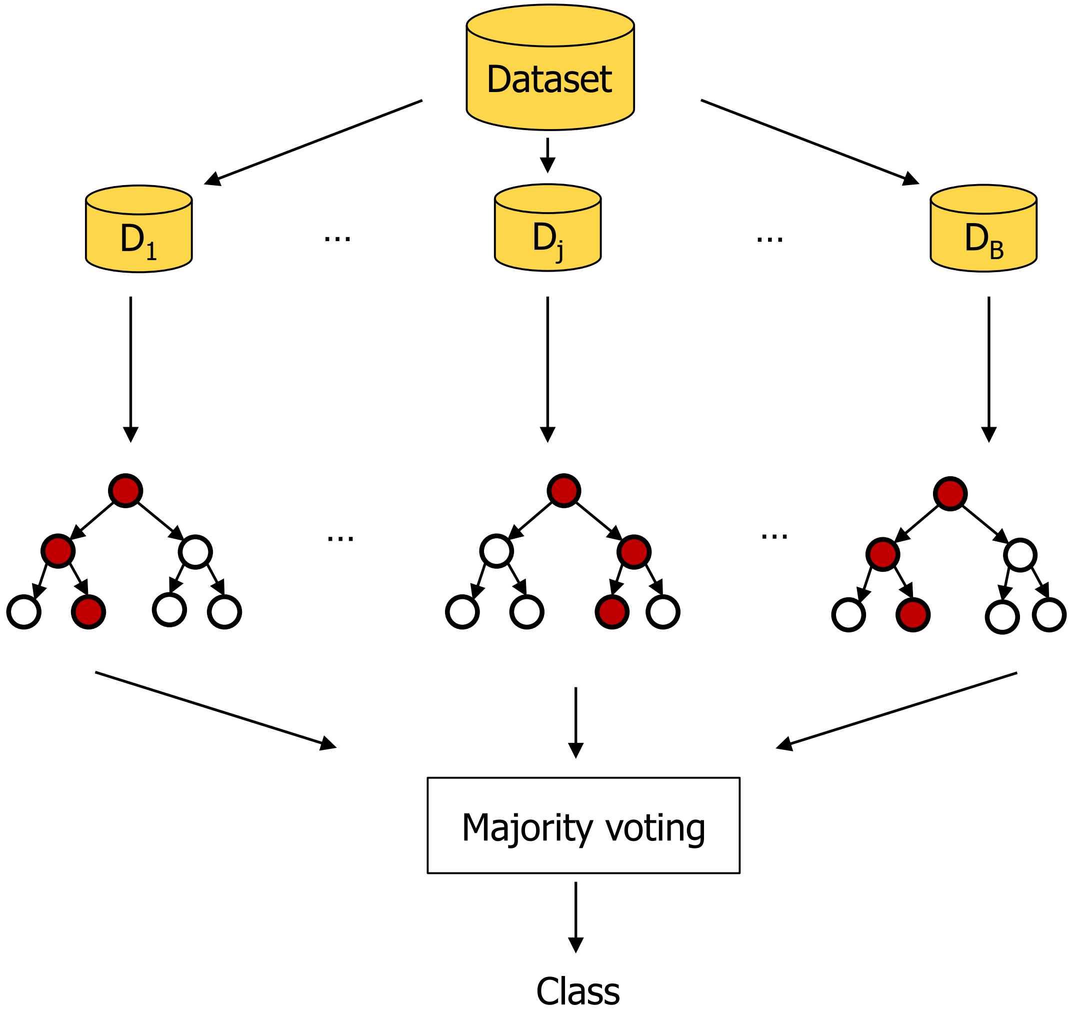
Original Training data

Random subsets

Multiple decision trees

For each subset, a tree is learned on a *random* set of features

Aggregating classifiers



Bootstrap aggregation

- Given a training set D of n instances, it selects B times a *random* sample with replacement from D and trains trees on these dataset samples
  - For b = 1, ..., B
    - Sample with replacement n' training examples, $n' \leq n$
      - A dataset subset D_b is generated
    - Train a classification tree on D_b

Feature Bagging

- Selects, for each candidate split in the learning process, a *random* subset of the features
  - being p the number of features, $\sqrt{p}$ features are typically selected
- Trees are decorrelated
  - Feature subsets are sampled randomly, hence different features could be selected as best attribute for the split

Random Forest – Algorithm Recap

- Given a training set D of n instances with p features
- For b = 1, ..., B
  - Sample randomly with replacement n' training examples. A subset D_b is generated
  - Train a classification tree on D_b
    - During the tree construction, for each candidate split
      - $m \ll p$ random features are selected (typically m ≈ $\sqrt{p}$)
      - the best split is computed among these m features
- Class is assigned by majority voting among the B predictions

Random Forest

- Strong points
  - higher accuracy than decision trees
  - fast training phase
  - robust to noise and outliers
  - provides global feature importance, i.e. an estimate of which features are important in the classification
- Weak points
  - results can be difficult to interpret
    - A prediction is given by hundreds of trees
      - but at least we have an indication through feature importance