Classifier Ensembles

Javier Béjar 🎓 🎓 🎓 🎓

LSI - FIB

Term 2012/2013
1. Classifier ensembles

2. Combining models

3. Application
1 Classifier ensembles

2 Combining models

3 Application
Weak classifiers

- One mayor drawback of decision trees/rule learners is that it is difficulty to approximate certain concepts
- Other drawback is its sensitivity to the sample of examples used for learning
- Slightly different samples could result in very different models (high variance)
- This means that the part of the error that we can not control (the variance) can be large
- For this reason decision trees are induction rule algorithms *weak classifiers*
Crowds are always right (?!)

Condorcet’s jury theorem

- Jury of voters who need to make a decision regarding a binary outcome
- Each voter has a probability $p$ of being correct
- Each voter is independent
- The probability of a majority of voters being correct is $L$
- Then
  - $p > 0.5$ implies $L > p$
  - $L$ approaches 1, for all $p > 0.5$ as the number of voters approaches infinity

This means that enough independent classifiers slightly better than random guessing can be predicted almost correctly
Crowds are sometimes right

- **Diversity of opinion**: Each member should have private information even if it is just an eccentric interpretation of the known facts
- **Independence**: Members’ opinions are not determined by the opinions of those around them
- **Decentralization**: Members are able to specialize and draw conclusions based on local knowledge
- **Aggregation**: Some mechanism exists for turning private judgments into a collective decision
This means that we can improve the accuracy of weak classifiers combining the decisions of a group of them (classifier ensembles).

An ensemble of simple models built using different datasets can obtain better accuracy than a more complex model.

Each model has a different “point of view” of the concept, their combination gives a more complete vision.

We are reducing the variance with their combination (and minimizing the error).
For decision trees, ensembles can be used also to obtain oblique decisions.

One decision tree

Many decision trees
1. Classifier ensembles

2. Combining models

3. Application
Classifier ensembles - Components

- **Training set**: A labeled dataset used for ensemble training
- **Base Inducer**: An induction algorithm that receives a training set and forms a classifier that represents the generalized relationship between the input attributes and the target attribute
- **Diversity Generator**: Responsible for generating the set of diverse classifiers
- **Combiner**: Responsible for combining the classifications of the ensemble
**Boosting:** Each instance in the dataset has a weight. Successive classifiers are generated increasing the weight of the instances that are not predicted correctly and decreasing the weight of the instances that are correctly predicted. Each classifier specializes on the difficult instances for the previous classifier.
Classifier ensembles - Adaboost Algorithm

**Algorithm:** Adaboost

**Input:** \( I \) (Weak inducer), \( T \) (iterations), \( E \) (training set, with binary labels \(+1,-1\))

**Output:** \( M_t \) (classifiers), \( \alpha_t \) (weights)

\[
t \leftarrow 1 \\
\text{foreach } i \in \{1..n\} \text{ do } D_1(i) = \frac{1}{|E|} \\
\text{repeat} \\
\quad \text{Build classifier } M_t \text{ using } E \text{ and distribution } D_t \\
\quad \epsilon_t \leftarrow \sum_{i:M_t(i) \neq y_i} D_t(i) \\
\quad \text{if } \epsilon_t > 0.5 \text{ then} \\
\quad \quad T \leftarrow t - 1 \\
\quad \quad \text{break} \\
\quad \text{end} \\
\quad \alpha_t \leftarrow \frac{1}{2} \log \frac{1-\epsilon_t}{\epsilon_t} \\
\quad \text{foreach } i \in \{1..n\} \text{ do } D_{t+1}(i) \leftarrow D_t(i) \cdot e^{-\alpha_t y_i M_t(i)} \\
\text{Normalize } D_t \text{ to a proper distribution} \\
\quad t++ \\
\text{until } t > T
Bagging (Bootstrapping aggregation): $n$ datasets are generated from an original dataset using extraction with replacement sampling. All classifiers are trained the same way.
Random Subspaces: We actually do not know what attributes are relevant for classifying the examples. We can select a subset of the attributes and obtain a more simple view of the task. Diversity is obtained using the attributes, not the examples. The same classifier trained in different subspaces can be combined to obtain a global view from different perspectives.

Example: Random forest, a combination of decision trees trained with subsets of attributes.
Classifier ensembles - Other approaches

**Voting:** Build N different classifiers (different algorithms) and use majority vote to obtain the result

**Stacking:** Build N different classifiers (different algorithms) to generate N new attributes and learn to classify the augmented dataset with other classifier

**Classification by regression:** Build N binary regression classifiers, one for each class, for each classifier assign +1 to the target class and -1 to the rest. The prediction is the classifier that obtains the higher positive value for its class.
1. Classifier ensembles

2. Combining models

3. Application
Heart disease diagnosis

13 Attributes (6 continuous, 1 discrete, 3 binaries, 4 qualitative)

Attributes: age, sex, chest pain type, resting blood pressure, serum cholestoral, fasting blood sugar, ...

270 instances

2 classes (present / not present)

Methods: Decision trees, induction rules, ensemble methods

Validation: 10 fold cross validation
Heart disease: Models

- DT Post pruning: accuracy 75.5% (35 nodes, 18 leaves)
- DT Unpruned: accuracy 76.3% (61 nodes, 31 leaves)
- DT Boosting: accuracy 81.5% (20 iterations, 8 inst leaf)
- DT Bagging: accuracy 82.2% (10 Prunned DT, 12 inst leaf)
- Induction rules: 76.6%
- IR Boosting: accuracy 80.4% (20 iterations)
- IR Bagging: accuracy 82.5% (40 IR, 0.6 IR sample)
- Random forest: accuracy 85.2% (5 random features, depth DT 3, 30 DT)
Heart disease: Decision tree
Heart disease: Decision rules

if chest <= 3.500 and age <= 56.500 then absent (9 / 77)
if thal > 4.500 and oldpeak > 0.550 and chest > 3.500 then present (57 / 2)
if sex <= 0.500 and age > 63.500 then absent (0 / 17)
if number_of_major_vessels <= 0.500 and serum_cholestoral <= 272
and maximum_heart_rate_achieved > 158.500 then absent (1 / 20)
if number_of_major_vessels > 0.500 and resting_electrocardiographic_results > 1
then present (24 / 2)
if resting_blood_pressure <= 122 and maximum_heart_rate_achieved <= 155
then absent (1 / 11)
if serum_cholestoral > 245.500 and serum_cholestoral <= 335 and oldpeak <= 0.100
then present (8 /1)
if resting_blood_pressure <= 127 and serum_cholestoral <= 324.500
then present (10 / 1)
if oldpeak <= 1.300 and age <= 58.500 then absent (0 / 9)
if maximum_heart_rate_achieved <= 127 and resting_blood_pressure > 129
then present (4 / 0)
else absent (5 / 9)