Cautious learning: Three-way out approach

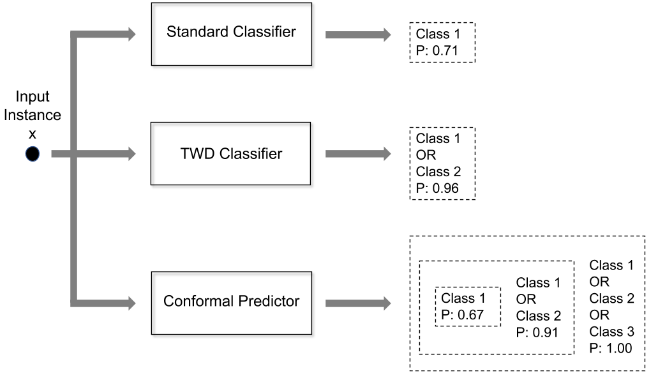
Davide Ciucci

Department of Informatics, Systems and Communication
University of Milan-Bicocca

Uncertainty in Computer Science

Cautious learning

- Not enough evidence to take a decision
- a generalization of supervised learning in which the Machine Learning (ML) models are allowed to express set-valued predictions



Three-way strategy

Define algorithms that can abstain

- a general method based on cost of abstention vs cost of error
- ad hoc methods: TW-decision tree, TW-random forest based on orthopartition

Result: three-way algorithms offer a trade-off among accuracy and coverage (the points that are classified)

TWO: decision theoretic approach

Given a probabilistic classifier transform it in a three-way classifier

TWO: decision theoretic approach

Given a probabilistic classifier transform it in a three-way classifier

STRATEGY 1 - ϵ AMBIGUITY - EXAMPLE

- Labels L $= \{1,2,3,4,5\}$

TWO: decision theoretic approach

Given a probabilistic classifier transform it in a three-way classifier

STRATEGY 1 - ϵ AMBIGUITY - EXAMPLE

- Labels L $= \{1,2,3,4,5\}$
- Probabilities for object x classification:
  $A(x) = \langle 0.2, 0.3, 0.15, 0.1, 0.25 \rangle$

TWO: decision theoretic approach

Given a probabilistic classifier transform it in a three-way classifier

STRATEGY 1 - ϵ AMBIGUITY - EXAMPLE

- Labels L = {1,2,3,4,5}
- Probabilities for object x classification: $A(x) = \langle 0.2, 0.3, 0.15, 0.1, 0.25 \rangle$
- Since $A(x)_2 = 0.3$ is the biggest, then the label of x is 2. However, 0.2 and 0.25 can be considered close to 0.3

TWO: decision theoretic approach

Given a probabilistic classifier transform it in a three-way classifier

STRATEGY 1 - ϵ AMBIGUITY - EXAMPLE

- Labels L $= \{1,2,3,4,5\}$
- Probabilities for object x classification:
  $A(x) = \langle 0.2, 0.3, 0.15, 0.1, 0.25 \rangle$
- Since $A(x)_2 = 0.3$ is the biggest, then the label of x is 2. However, 0.2 and 0.25 can be considered close to 0.3
- The classification of x is ambiguous: $\{1,2,5\}$

TWO: decision theoretic approach

Given a probabilistic classifier transform it in a three-way classifier

STRATEGY 2: balance the cost of errors and abstention

TWO: decision theoretic approach

Given a probabilistic classifier transform it in a three-way classifier

STRATEGY 2: balance the cost of errors and abstention

1. set a cost of error and abstention
2. define the risk of a decision (using probabilities A(x))
3. the decision is the less risky set of labels

TWO: decision theoretic approach

Some more details:

- ϵ cost of prediction error
  If the error cost is constant, the complexity of the procedure is $O(n)$

TWO: decision theoretic approach

Some more details:

- ϵ cost of prediction error
  If the error cost is constant, the complexity of the procedure is $O(n)$
- $\alpha : \mathcal{P}(X) \mapsto \mathfrak{R}$ cost of partial abstention
  $\alpha(Z)$ the cost of abstaining among the alternatives in Z

TWO: decision theoretic approach

Some more details:

- ϵ cost of prediction error
  If the error cost is constant, the complexity of the procedure is $O(n)$
- $\alpha : \mathcal{P}(X) \mapsto \mathfrak{R}$ cost of partial abstention
  $\alpha(Z)$ the cost of abstaining among the alternatives in Z
- The risk of decision Z

$$R(Z) = \alpha(Z) \cdot \sum_{y_i \in Z} A(x)_i + \epsilon \sum_{y_j \notin Z} A(x)_j$$

TWO: decision theoretic approach

STRATEGY 2 - EXAMPLE

- Labels L $= \{1,2,3,4,5\}$
- Probabilities for object x classification:
  $A(x) = \langle 0.2, 0.3, 0.15, 0.1, 0.25 \rangle$

TWO: decision theoretic approach

STRATEGY 2 - EXAMPLE

- Labels L $= \{1,2,3,4,5\}$
- Probabilities for object x classification: $A(x) = \langle 0.2, 0.3, 0.15, 0.1, 0.25 \rangle$
- $\epsilon = 1$ and $\alpha(Z) = \frac{|Z|-1}{|Y|-1}$

TWO: decision theoretic approach

STRATEGY 2 - EXAMPLE

- Labels L $= \{1,2,3,4,5\}$
- Probabilities for object x classification: $A(x) = \langle 0.2, 0.3, 0.15, 0.1, 0.25\rangle$
- $\epsilon = 1$ and $\alpha(Z) = \frac{|Z|-1}{|Y|-1}$
- Compute the risk for all sets containing label '2'
  - $R(\{1,2\}) = \frac{1}{4}(0.2 + 0.3) + 1(0.15 + 0.1 + 0.25) = 0.625$

TWO: decision theoretic approach

STRATEGY 2 - EXAMPLE

- Labels L = $\{1,2,3,4,5\}$
- Probabilities for object x classification:
  $A(x) = \langle 0.2, 0.3, 0.15, 0.1, 0.25 \rangle$
- $\epsilon = 1$ and $\alpha(Z) = \frac{|Z|-1}{|Y|-1}$
- Compute the risk for all sets containing label '2'
  - $R(\{1,2\}) = \frac{1}{4}(0.2 + 0.3) + 1(0.15 + 0.1 + 0.25) = 0.625$
  - $R(\{2\}) = 0 \cdot 0.3 + 1 \cdot 0.7$
  - $\ldots\ldots$
- The less risky is $Z = \{2, 5\}$

Model specific strategies

The previous strategies are

- generic: take the results of a probabilistic classifier and transform it in a three-way
- do not exploit directly the ambiguity in data

Model specific strategies

The previous strategies are

- generic: take the results of a probabilistic classifier and transform it in a three-way
- do not exploit directly the ambiguity in data

We implemented modifications of

- Decision Trees
- Random Forest
- Optimization Based Learning (logistic regression, Support Vector Machines, etc.)

Model specific strategies

The previous strategies are

- generic: take the results of a probabilistic classifier and transform it in a three-way
- do not exploit directly the ambiguity in data

We implemented modifications of

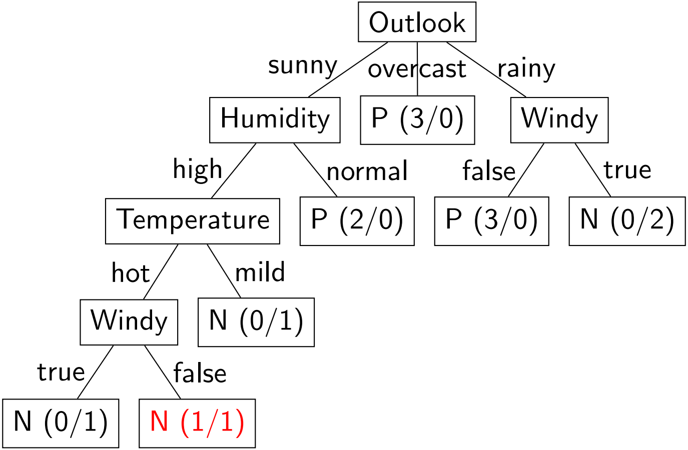
- Decision Trees
- Random Forest
- Optimization Based Learning (logistic regression, Support Vector Machines, etc.)

More details on Decision Trees

Decision Tree

Temperature	Outlook	Humidity	Windy	Do Sport?
hot	sunny	high	false	no
hot	sunny	high	true	no
hot	sunny	high	false	yes
cool	rain	normal	false	yes
cool	overcast	normal	true	yes
mild	sunny	high	false	no
cool	sunny	normal	false	yes
mild	rain	normal	false	yes
mild	sunny	normal	true	yes
mild	overcast	high	true	yes
hot	overcast	normal	false	yes
mild	rain	high	true	no
cool	rain	normal	true	no
mild	rain	normal	false	yes

Decision Tree (ID3)



The idea

A classification can be

$$\text{YES}/\text{NO}/\text{UNDECIDED}$$

Two steps

1. *Define an orthopartition from each attribute*
2. Select as split attribute the one with greatest mutual information wrt the decision

From an attribute to an orthopartition

When to abstain from a decision? When it is less costly!

From an attribute to an orthopartition

When to abstain from a decision? When it is less costly!

- Two parameters $\alpha < \epsilon$ to weight errors
  - α the cost of an abstention
  - ϵ: the cost of a classification error

From an attribute to an orthopartition

When to abstain from a decision? When it is less costly!

- Two parameters $\alpha < \epsilon$ to weight errors
  - α the cost of an abstention
  - ϵ: the cost of a classification error
- Compute total error for each attribute *a* and each value *i*
- If total classification error $\geq$ total abstention error $\rightarrow$ better to abstain

From an attribute to an orthopartition

- For an attribute a, for each value we assign a decision: yes/no/$\perp$

From an attribute to an orthopartition

- For an attribute a, for each value we assign a decision: yes/no/$\perp$
  - D_i^a objects with value i: $D_i^a = \{x \in D | v_a(x) = i\}$

From an attribute to an orthopartition

- For an attribute a, for each value we assign a decision: yes/no/$\perp$
  - D_i^a objects with value i: $D_i^a = \{x \in D | v_a(x) = i\}$
  - The elements in D_i^a are in majority classified as yes or no?

    We associate to D_i^a the classification
    $C_i^a = argmax_{j \in \{yes, no\}}\{|\{x \in D_i^a | C(x) = j\}|\}$

From an attribute to an orthopartition

- For an attribute a, for each value we assign a decision: yes/no/$\perp$
  - D_i^a objects with value i: $D_i^a = \{x \in D | v_a(x) = i\}$
  - The elements in D_i^a are in majority classified as yes or no?

    We associate to D_i^a the classification
    $C_i^a = argmax_{j \in \{yes,no\}}\{|\{x \in D_i^a | C(x) = j\}|\}$
- and compute the error/abstention costs
  - Expected classification error cost

$$E(D_i^a | C_i^a) = \epsilon * min_{j \in \{yes,no\}}\{|\{x \in D_i^a | C(x) = j\}|\}$$

From an attribute to an orthopartition

- For an attribute a, for each value we assign a decision: yes/no/$\perp$
  - D_i^a objects with value i: $D_i^a = \{x \in D | v_a(x) = i\}$
  - The elements in D_i^a are in majority classified as yes or no?

    We associate to D_i^a the classification
    $C_i^a = argmax_{j \in \{yes,no\}}\{|\{x \in D_i^a | C(x) = j\}|\}$
- and compute the error/abstention costs
  - Expected classification error cost

$$E(D_i^a | C_i^a) = \epsilon * min_{j \in \{yes,no\}}\{|\{x \in D_i^a | C(x) = j\}|\}$$

  - Expected abstention error cost

$$E(D_i^a | \perp) = \alpha |D_i^a|$$

From an attribute to an orthopartition

- If $E(D_i^a|C_i^a) < E(D_i^a|\perp)$ we assign to the objects in D_i^a the decision C_i^a otherwise, the decision is $\perp$

From an attribute to an orthopartition

- If $E(D_i^a|C_i^a) < E(D_i^a|\perp)$ we assign to the objects in D_i^a the decision C_i^a otherwise, the decision is $\perp$
- Union over all values $i \rightarrow$ define an orthopair $O_a = (P_a, N_a)$

$$P_a = \bigcup \{D_i^a | C_i^a = yes\} \text{ and } N_a = \bigcup \{D_i^a | C_i^a = no\}$$

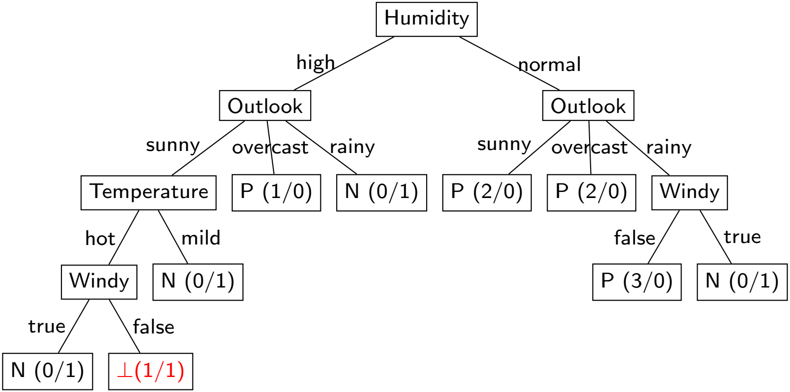
- Define the orthopartition $\mathcal{O}_a = \{O_a, \neg O_a\}$

The algorithm

Input: Dataset D, error cost ϵ, abstention cost α
Output: Three-way Decision Tree built on D

1 Feature $a \rightarrow$ orthopartition $\mathcal{O}_a$ using ϵ, α;
2 Orthopartition $\mathcal{O}_a \rightarrow$ mutual information $m(D, \mathcal{O}_a)$;
3 split attribute $=$ the feature a_{max} which gives the greatest mutual information value;
4 Recur on the subsets of D determined by a_{max};

Example



Humidity
high
normal
Outlook
Outlook
sunny
overcast
rainy
sunny
overcast
rainy
Temperature
P (1/0)
N (0/1)
P (2/0)
P (2/0)
Windy
hot
mild
false
true
Windy
N (0/1)
P (3/0)
N (0/1)
true
false
N (0/1)
⊥(1/1)

Some comments

- Not discussed here
  - Extension of the method to more than two-valued (yes/no) decisions

Some comments

- Not discussed here
  - Extension of the method to more than two-valued (yes/no) decisions
- In case of indecision the algorithm returns a subset of decisions: the correct one is always included in this subset

Some comments

- Not discussed here
  - Extension of the method to more than two-valued (yes/no) decisions
- In case of indecision the algorithm returns a subset of decisions: the correct one is always included in this subset
- problem: accuracy depends on arbitrary error weights ϵ and α

TWO experiments

- Compared KNN,Logistic Regression, Random Forest, Naive Bayes, SVM and their 3-way variants

TWO experiments

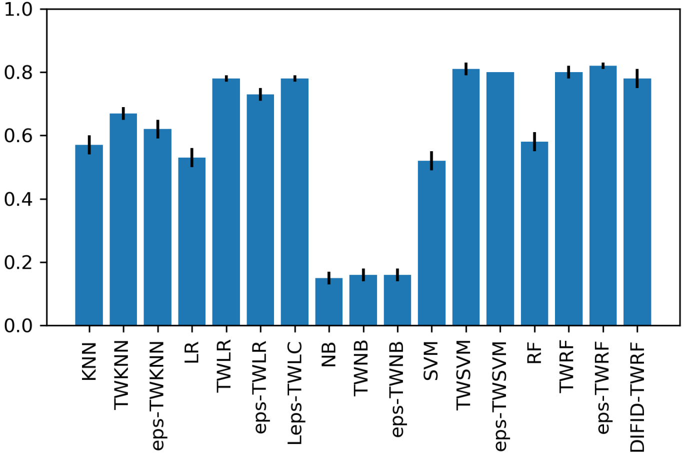
- Compared KNN,Logistic Regression, Random Forest, Naive Bayes, SVM and their 3-way variants
- 6 UCI datasets + 1 real-world medical dataset

Dataset	# instances	# attributes	# classes
Iris	150	4	3
Wine	178	13	3
Digits	1797	64	10
Breast cancer	569	30	2
Olivetti faces	400	4096	40
Yeast	1484	8	10
SF12	462	10	2

TWO experiments

The three-way versions (Strategies 1/2) are better than the standard version

Yeast dataset



TWO experiments

The best algorithms are the ones derived from random forest

Alg.	TWRF	DIFID-TWRF	ϵ-TWRF	L_ϵ-TWLC	TWLR	TWSVM	TWKNN	RF	KNN/ϵ-TWLR/ϵ-TWSVM
Rank	2.14	2.28	2.71	3.71	4.00	4.14	4.42	4.86	5.86

Table: Average ranks of the top 10 performing algorithms.

TWO experiments

The best algorithms are the ones derived from random forest

Alg.	TWRF	DIFID-TWRF	ϵ-TWRF	L_ϵ-TWLC	TWLR	TWSVM	TWKNN	RF	KNN/ϵ-TWLR/ϵ-TWSVM
Rank	2.14	2.28	2.71	3.71	4.00	4.14	4.42	4.86	5.86

Table: Average ranks of the top 10 performing algorithms.

- No significant differences among strategy 1, strategy 2 and ad-hoc algorithms
- Strategy 1: comparable performance but with less parameters to set and increased computational efficiency

Conclusions

- The capability of directly using and conveniently communicating the ambiguity encountered by the algorithm in recommending a class could be critical to deliver reliable Machine Learning-based Decision Support Systems

Conclusions

- The capability of directly using and conveniently communicating the ambiguity encountered by the algorithm in recommending a class could be critical to deliver reliable Machine Learning-based Decision Support Systems
- Abstention in ML output is a way to trade (decision) accuracy with efficiency: unresolved advice implies that decision-makers have to look for and consider more evidence, even beyond the available data