Three–Way Classification: Ambiguity and Abstention in Machine Learning

Andrea Campagner[1,3], Federico Cabitza[1,2], and Davide Ciucci[1](✉)

[1] Dipartimento di Informatica, Sistemistica e Comunicazione,
University of Milano–Bicocca, Viale Sarca 336, 20126 Milan, Italy
davide.ciucci@unimib.it

[2] IRCCS Istituto Ortopedico Galeazzi, Via Galeazzi 4, 20161 Milan, Italy

[3] Deloitte Italia, Via Tortona 25, Milan, Italy

Abstract. Ambiguity, that is the lack of information to produce a specific classification, is an important issue in decision–making and supervised classification. In case of ambiguity, human–decision makers can resort to abstaining from making precise classifications (especially when error-related costs are high), but this behaviour has been scarcely addressed, and applied, in machine learning algorithms. This contribution grounds on previous works in the areas of three–way decisions, cautious classification and orthopairs, and proposes a set of techniques we developed to address this form of ambiguity, by providing both a general–purpose technique to create three–way algorithms from probabilistic ones, and also more specific techniques which could be applied to popular machine learning frameworks. We also evaluate the proposed idea, by performing a set of experiments where we compare classical classification algorithms with the corresponding three–way generalizations, in order to study the trade–off between classification accuracy and abstention: the results are promising.

Keywords: Machine Learning · Abstention · Three–way decision · Data Mining · Ambiguity · Orthopair · Orthopartition · Uncertainty

1 Introduction

Research in the *Machine Learning* and *Data Mining* fields has recently taken central stage in the Computer Science research community: this interest has been driven by *theoretical* advancements [3,11,17,23], *technological* advancements and, chiefly among all, the promising results in different application areas (driven by the availability of large amounts of data) [13,24,26].

Despite all the attention and recent achievements, a limitation of current Machine Learning methods is the inability to properly deal with uncertainty and biases affecting the training datasets which are fed to learning algorithms as input [15,18,30]. Indeed, as noted in [4] for the healthcare domain, various forms of uncertainties and biases can affect the training data (*missing data*,

T. Mihálydeák et al. (Eds.): IJCRS 2019, LNAI 11499, pp. 280–294, 2019.
https://doi.org/10.1007/978-3-030-22815-6_22

inter–rater disagreement, *lack of information*, *ambiguity*, ...) thus hampering the performance and, most relevantly, the *reliability* of the resulting models.

Several uncertainty theories (e.g. *rough–set theory* [27], *fuzzy–set theory* [32], *three–way decisions* [31],...) have been proposed in order to cope with these different forms of uncertainty, also with application to Machine Learning [2,21], but their adoption in mainstream Machine Learning has been lagging, due to different reasons (e.g. those evidenced in [22] for the case of fuzzy sets).

In this work, we will consider a particular type of uncertainty, *lack of information*, also called *ambiguity* in the terminology of [19]. In the decision–making/classification domain, this type of uncertainty occurs when a human (or computational) agent deems the available information insufficient to cast a univocal and reasonable decision.

Whenever possible, the usual strategy that human decision–makers adopt, in order to cope with either ambiguous input or uncertain output, is to reject any pretense of giving a *clear–cut* decision and, instead, *abstain* from expressing a judgment. This approach has the merit of highlighting, in a simple form, which instances are more *uncertain* and, consequently, pointing out which ones would require the acquisition of further information.

While this approach is still little adopted, different authors have tried to address the *abstention* behaviour under a computational perspective: here, we especially mention the work on *cautious classifiers* [16,20] and the work on three–way decisions [31]. In the same direction, in order to develop Machine Learning models with this *abstention* ability, the authors proposed in [5,6] an extended *decision tree–learning* model, based on *orthopairs* [9,10] and three–way decisions.

In this article, we extend this line of research:

- We introduce a general framework for classification with abstention (or three–way classification), based on three–way decisions and orthopartitions, which can be applied to any classification algorithm;
- We define a set of specific strategies which can be used to directly implement three–way classification in the context of popular learning algorithms (e.g. decision trees, random forests, logistic regression);
- We conduct an experimental study, in which we compare different classical learning algorithms with the corresponding three–way ones on various datasets.

More specifically, in Sect. 2, we give a basic introduction to orthopairs and orthopartitions. In Sect. 3 the basic methods are introduced, that is: in Sect. 3.1, we define our approach to convert any classifier into a three–way classification algorithm, both in the binary and multi–class settings, providing also a theoretical–algorithmic analysis of these frameworks; in Sects. 3.2 and 3.3, we describe the strategies to directly implement three–way classification for three popular learning models (i.e., Decision Trees, Random Forests and Convex Learning via Gradient Descent). In Sect. 4.1, we illustrate the setting of the empirical analysis we conducted in order to compare traditional learning algorithms with three–way ones. In Sect. 4.2, we present the results of the conducted experiments, considering the advantages offered by three–way classification algorithms and evaluating the effect of abstention with respect to their performance, supporting our analysis with standard statistical validation techniques. Finally, in Sect. 5 we present our conclusions and outline the set of open problems and issues that we plan to investigate in our future works.

2 Orthopairs and Orthopartitions

Let us recall some basic notions on orthopairs and orthopartitions [6,10].

Let U be a set of objects, an orthopair is a pair $O = \langle P, N \rangle$ of subsets of U such that $P \cap N = \emptyset$. From these two sets we can also define the *boundary* as $Bnd = (P \cup N)^c$. Note that we could take an orthopair as a partially specified set which expresses our (incomplete) knowledge about the assignment of objects in a universe to a certain concept class; in this case, set P represents the positive examples for the concept while N represents the negative ones. We say that a set S is consistent with an orthopair O if it holds that:

$$x \in P \rightarrow x \in S \wedge x \in N \rightarrow x \notin S$$

That is, if we interpret the orthopair O as a partially specified set expressing our degree of knowledge about the belonging (or not) of certain objects to a set, S is coherent with our partial knowledge.

We say that two orthopairs O_1, O_2 are *disjoint* if it holds that:

(Ax D1) $P_1 \cap P_2 = \emptyset$;
(Ax D2) $P_1 \cap Bnd_2 = \emptyset$ and $Bnd_1 \cap P_2 = \emptyset$.

Definition 1. *An* orthopartition *is a set $\mathcal{O} = \{O_1, ..., O_n\}$ of orthopairs such that the following axioms hold:*

(Ax O1) $\forall O_i, O_j \in \mathcal{O}$ O_i, O_j are disjoint;
(Ax O2) $\bigcup_i (P_i \cup Bnd_i) = U$;
(Ax O3) $\forall x \in U$ $(\exists O_i$ s.t. $x \in Bnd_i) \rightarrow (\exists O_j$ with $i \neq j$ s.t. $x \in Bnd_j)$;
(Ax O4) $|\mathcal{O}| \leq |U|$

It can be observed that an orthopartition represents a *partial classification*, or a *classification with abstentions* (in a multi–class setting): the objects in the boundaries represent those objects whose class assignment is not precisely known (given the available evidence and, hence, the presence of ambiguity).

Definition 2. *A partition π is* consistent with an orthopartition $\mathcal{O}$ *iff $\forall O_i \in \mathcal{O}$, $\exists! \, S_i \in \pi$ such that S_i is consistent with O_i. We denote with $\Pi_{\mathcal{O}}$ the set of all partitions consistent with $\mathcal{O}$: $\Pi_{\mathcal{O}} = \{\pi | \pi$ is consistent with $\mathcal{O}\}$.*

Viewing an orthopartition as a partial state of knowledge about a multi–class classification (associated with the set $\Pi_{\mathcal{O}}$ which represents all possible consistent complete states of knowledge), we can extend many measures defined on classical partitions to orthopartitions, in particular we will focus on the *entropy* and *accuracy* (the extension of other metrics based on the confusion matrix is analogous). The *logical entropy* [14] of a partition π is defined as:

$$h(\pi) = \tfrac{dit(\pi)}{|U|^2}$$

where $dit(\pi) = \{(u, u') \in U \times U | \, u, u' \text{ belong to two different blocks of } \pi\}$. We can define three different generalizations of this concept, when applied to orthopartitions:

Definition 3. *Given an orthopartition $\mathcal{O}$, we define the* lower entropy*, the* upper entropy *and the* mean entropy *respectively as:*

$$h_* = min\{h(\pi) | \pi \in \Pi_{\mathcal{O}}\} \tag{1a}$$
$$h^* = max\{h(\pi) | \pi \in \Pi_{\mathcal{O}}\} \tag{1b}$$
$$h_A = \frac{1}{|\Pi_{\mathcal{O}}|} \sum_{\pi \in \Pi_{\mathcal{O}}} h(\pi) \tag{1c}$$

As shown in [6,7], all three values can be computed in polynomial time. Let π_1, π_2 be two partitions and $f : \pi_1 \mapsto \pi_2$ be a bijection between the blocks of π_1, π_2, the accuracy of π_2 wrt π_1 is defined as:

$$acc_{\pi_1}(\pi_2) = \frac{1}{|U|} \sum_{S_i \in \pi_1} |S_i \cap f(S_i)|$$

Similarly, we can provide three generalizations of the accuracy:

Definition 4. *Given a partition π^*, an orthopartition $\mathcal{O}$, and a bijection f between the respective blocks, we define the* lower accuracy*, the* upper accuracy *and the* mean accuracy *respectively as:*

$$acc_* = min\{acc(\pi) | \pi \in \Pi_{\mathcal{O}}\} \tag{2a}$$
$$acc^* = max\{acc(\pi) | \pi \in \Pi_{\mathcal{O}}\} \tag{2b}$$
$$acc_A = \frac{1}{|\Pi_{\mathcal{O}}|} \sum_{\pi \in \Pi_{\mathcal{O}}} acc(\pi) \tag{2c}$$

Another interesting measure of accuracy (that we denote as acc_O) is obtained by considering, in the computation of the accuracy value, only the instances which are not in the boundary regions: that is, if $U_r \subseteq U$ is the restriction of U to the objects which are not placed in boundaries for orthopartition $\mathcal{O}$ then:

$$acc_O = \frac{1}{U_r} \sum_{S_i \in \pi_1} |S_i \cap f(S_i)|$$

where S_i and $f(S_i)$ are similarly restricted to U_r.

3 The Methods

In this section, we propose the main method of three-way classification and apply it to different learning strategies.

3.1 Three–Way Classification

Let $Y = \{y_1, ..., y_k\}$ be a set of class labels, $X = \{x^1, ..., x^n\}$ be a set of objects, $C : X \rightarrow Y$ be a function which associates with each object $x^i \in X$ its true classification $y^i_j \in Y$. Let A be a probabilistic classifier, that is, an algorithm which, given an object $x^i \in X$, returns a probability distribution $A(x^i)$ over Y, that is, $A : X \rightarrow \mathcal{P}(Y)$, where $\mathcal{P}(Y)$ is the space of probability distributions over Y. For each $y_j \in Y$, $A(x^i)_j$ represents the probability that algorithm A assigns to the event that y_j is the correct class labeling for object x^i (i.e., the subjective probability that $C(x^i) = y_j$). Typically, in the Machine Learning domain, this *soft* probabilistic classification is then converted into an *hard* one by selecting the $y_j \in Y$ with maximum probability: that is, we define $D(x^i) = argmax_{y_j \in Y} A(x^i)_j$ and we denote with $A(x^i)^*$ the corresponding probability. Note that this classification rule completely hides away the uncertainty of the classifier and, consequently, the ambiguity intrinsic in its input. An approach to let the classifier A fully express its uncertainty, which fully reflects the ambiguity of its input datum, is to let the classifier *abstain* on those instances whose assignment to the classification labels is considered ambiguous.

First, we limit ourselves to a binary classification problem, that is, $Y = \{0, 1\}$. Let ϵ be the cost associated with an erroneous classification, and let τ the cost associated with an abstention. Let $x \in X$ be an object, it is evident and widely known [8,16,31] that, in this context, algorithm A should choose to abstain on x if:

$$\tau < \epsilon * min_{j \in \{0,1\}} A(x^i)_j$$

that is, if choosing to abstain would incur (in the expected value) a lower cost than adopting a clear-cut classification (selected using the standard decision rule). The same decision rule could be given using a probability threshold; it is easy to show that the two formulations are equivalent.

Theorem 1. *Algorithm A should select to abstain iff* $max_{j \in \{0,1\}} A(x^i)_j < 1 - \frac{\tau}{\epsilon}$

Proof. Let $A(x)^* = max_{j \in \{0,1\}} A(x^i)_j$, the rule expressed above is equivalent to $\tau < \epsilon * (1 - A(x)^*) \Rightarrow \frac{\tau}{\epsilon} < 1 - A(x)^* \Rightarrow A(x)^* < 1 - \frac{\tau}{\epsilon}$.

The generalization to the multi–class setting, in which partial decisions could also be expressed, is also feasible and clearly more interesting. Indeed, in [6], a generalization of this classification rule is proposed as follows. Let $Z \subseteq Y$, then in this context we allow the algorithm A to express a decision Z, by which we mean that the algorithm is confident that the true label of x is in Z but it is unsure about its precise identity. Let $A(x)_Z = \sum_{y_j \in Z} A(x)_j$. If, as in the binary classification setting, we adopt a constant abstention cost τ, then the algorithm, with the abstention decision rule, should abstain on instance x if:

$$\tau * A(x)_{Z^*} + \epsilon * A(x)_{Y \setminus Z^*} < \epsilon * (1 - A(x)^*) \tag{3}$$

where $Z^* \subseteq Y$ is the set of labels which minimizes the left hand of the inequality, otherwise it should output the y_j corresponding to $A(x)*$.

Note that, directly translating this definition (as done in [6]) to an algorithm, yields a decision procedure which has complexity *exponential* w.r.t. $|Y|$. However, it is easy to observe that not every $Z \subseteq Y$ should be considered in the above minimization problem. In fact, the above minimization problem can be solved correctly in a *greedy* approach: let $\overset{\wedge}{A(x)} = \langle y_1^*, ..., y_k^* \rangle$ be the result of sorting $A(x)$ in order of decreasing probability. Then the above decision rule can be expressed, without loss of generalization, as:

$$\tau * \sum_{i=1}^{j} \overset{\wedge}{A(x)}_i + \epsilon * \sum_{i=j+1}^{k} \overset{\wedge}{A(x)}_i < \epsilon * (1 - A(x)^*) \tag{4}$$

where j is the index which minimizes the left hand of the inequality.

Theorem 2. *The greedy version of the optimization algorithm is solvable with time complexity $\Theta(n)$ (if $A(x)$ is already sorted).*

Proof. For each j we can pre-compute $\sum_{i=1}^{j} \overset{\wedge}{A(x)}_i$ in constant time (by accumulating the values of the sum over previous js), from this value we can obtain $\sum_{i=j+1}^{k} \overset{\wedge}{A(x)}_i$ in constant time. The result easily follows.

As observed in [6], a constant value of τ has the result that, when the algorithm abstains, Z^* (i.e. the set of labels which minimizes the optimization problem) is always $Z^* = Y$. This problem can be solved in a *regularization* fashion, by penalizing overly uncertain responses from the algorithm. In this case τ is defined as a function $\tau : \{1, ..., |Y|\} \to \mathbb{R}_+$ such that, given $A, B \subseteq Y$, it holds $|A| \leq |B| \to \tau(|A|) \leq \tau(|B|)$.

An interesting aspect to note is that not every value of τ is meaningful in this context, namely the following result holds:

Theorem 3. *Let us consider a* n–class *classification problem. Abstention can be achieved only if $\tau < \epsilon * \frac{n-1}{n}$.*

Proof. Consider the case of constant τ and the formulation given by Eq. 3. Then, we have that the algorithm should decide to abstain iff $\tau < \epsilon * (1 - A(x)^*)$. But $A(x)^* \geq \frac{1}{n}$, thus $\tau < \epsilon * (1 - A(x)^*) \leq \epsilon * (1 - \frac{1}{n})$, from which we obtain the result.

Example 1. Let x be an instance and A a probabilistic algorithm, defined over the label set $Y = \{1, 2, 3, 4\}$ such that $A(x) = \begin{bmatrix} 0.3 \; 0.3 \; 0.2 \; 0.2 \end{bmatrix}$, and let $\epsilon = 1$, $\tau = 0.4$. Then, the right hand of Eq. (3) is 0.7, while Z^* can be verified to be, as expected, $Z^* = Y$ with the left hand of inequality (3) assuming value 0.4. If, on the other hand, we do not assume a constant τ but instead adopt $\tau(Z) = 0.4 \cdot \frac{1}{1-\frac{|Z|-2}{|Y|}}$, thus penalizing abstentions over a larger set of alternatives, we have that $Z^* = \{1, 2, 3\}$ (equivalently, $Z^* = \{1, 2, 4\}$) and the left hand of the inequality has value 0.63.

3.2 Decision Trees and Random Forests

In [6] an extended Decision Tree model, called Three–Way Decision Tree (TWDT), is proposed. It provides a more tight integration of Decision Trees and Three–Way Classification than the main approach described in this paper. Let $D = \{x_1, ..., x_{|D|}\} \subseteq X$ be a given dataset with a set of features $\{a_1, ..., a_m\}$. We denote by $D_i^a = \{x \in D | v_a(x) = v_i^a\}$ the set of instances that have value v_i^a for feature a. We associate with D_i^a the classification C_i^a, which is obtained by the decision rule described in Sect. 3.1 (note that this class assignment is done locally on the tree nodes, and not only on the final output of the classifier). Since this classification determines an orthopartition $\mathcal{O}_a$, we can then compute the *accuracy* of $\mathcal{O}_a$ w.r.t. D as described in Sect. 2 (selecting among acc_*, acc^*, acc_A) and choose the feature a^* which results in the maximum accuracy value, and then recur (until a termination criterion is met) on the subsets of D determined by feature a^*.

This approach can be easily extended to Random Forests (or other ensemble learning algorithms). Basically, the learning process, as in standard Random Forest learning, first induces a set of n TWDT estimators, which we denote as $T_1, ..., T_n$. Each of these TWDT estimators can be viewed as an orthopartition $\mathcal{O}_i = \{\langle P_{y_1}, N_{y_1}\rangle, ...\langle P_{y_k}, N_{y_k}\rangle\}$ on the set of instances X, which assigns a set of labels $T_i(x) \subseteq Y$ to each instance $x \in X$.

Let $x \in X$ be a new instance to classify, then the ensemble of trees $T_1, ..., T_n$ determines a *basic belief assignment* (BBA) (in the sense of *evidence theory* [28]) $m_x(S) = \frac{|\{T_i | T_i(x) = S\}|}{n}$. This BBA could then be transformed to a probability distribution using the *pignistic transformation* [29] $p(y_j) = \sum_{S \ni y_j} \frac{m(S)}{|S|}$, obtaining a probabilistic classifier to which the decision procedure described in Sect. 3.1 could be applied.

3.3 Convex Learning Approximation

Several ML approaches (e.g. logistic regression, SVMs, multi–layer neural networks, ...) are based on the Gradient Descent algorithm, which is used to iteratively update the parameters of the models by taking in consideration the gradient of a loss function w.r.t. the parameters. A caveat, in order to ensure that the algorithm converges to a global minimum, is that the loss function should be a convex function. It is easy to note that the decision rule described in Sect. 3.1 (which could be seen as a generalized version of the standard *0–1 loss*) does not result in a convex loss function:

Theorem 4. *The loss function determined by the decision procedure described in Sect. 3.1 is not convex.*

Proof. Let $D(x^i) = \begin{cases} Z* & \exists Z^* \text{which solves Eq. 3,} \\ \overset{\wedge}{y^i} & otherwise \end{cases}$

Then the loss of algorithm A w.r.t to instance x is $L(x) = \begin{cases} 0 & D(x) = C(x) \\ \tau & C(x) \in D(x) \\ \epsilon & otherwise \end{cases}$

Clearly, $L(x)$ is not convex.

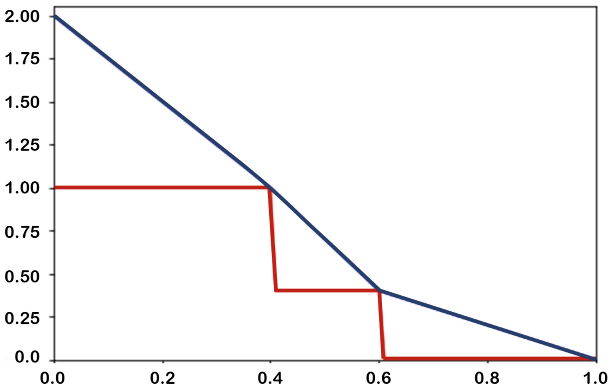

Fig. 1. The depiction of the loss function (in red), and its convex piece–wise linear approximation (in blue), for positive examples. (Color figure online)

We can, however, define a convex approximation of the above described loss function [1]. Consider first a binary classification problem, the loss function described above is depicted in Fig. 1. As shown in Fig. 1 we can, however, define a convex piece–wise linear approximation to the real loss. Consider first a binary classification problem assuming, without loss of generality, that $\epsilon = 1$. For the positive examples (i.e., those $x \in X$ s.t. $C(x) = 1$) we can express an approximation *from above* (so that we never underestimate the error) as:

$$L(w) = max\{0, 1 - w, \frac{(2 * \tau - 1) * w + 3 * \tau - 1 - \tau^2}{2 * \tau - 1}, 2 - \frac{w}{\tau}\} \tag{5}$$

where $w = A(x)_1$ (i.e. the probability that algorithm A assigns object x to the positive class).

Theorem 5. *The loss function described in Eq. 5 is convex.*

Proof. Each of the arguments of the *max* function is linear in w, thus it is convex (every linear function is both convex and concave). Furthermore the point–wise *max* of convex function is convex, from which the statement follows.

The expression for negative examples is equivalent and symmetric. This loss function could be then used to directly train convex learning algorithms and, given a new instance $x \in X$ to classify, we compute $A(x)_1$ and then, we classify x using the decision rule defined in Sect. 3.1.

In order to extend this approach to multi–class classification, we simply adopt a *one–vs–one* learning scheme, in which, for each pair of labels $y_i, y_j \in Y$ we train a classifier $A^{i,j}$ using the convex loss function described above. Then, given a new instance $x \in X$ to classify, we compute for each classifier $A^{i,j}$ its output $D^{i,j}(x)$ and we implement a voting schema:

$$vote_y(x, A^{i,j}) = \begin{cases} 1 & D^{i,j}(x) = y \\ \frac{1}{|D^{i,j}(x)|} & y \in D^{i,j}(x) \\ 0 & otherwise \end{cases}$$

and the final votes are computed as $vote_y(x) = \sum_{A^{i,j}} vote_y(x, A^{i,j})$ which, again, determines a probabilistic classifier to which the decision rule described in Sect. 3.1 can be applied.

4 Experimental Comparison

In order to test the flexibility offered by allowing an abstention decision (or a set of abstention decisions, in the multi–class setting) we performed an experimental comparison, analyzing a variety of traditional ML algorithms and their respective Three–Way generalization, on a set of datasets. More specifically, we considered the following algorithms: *k–nearest neighbors* (KNN), *logistic regression* (LR), *linear discriminant analysis* (LDA), *Naive Bayes* (NB), *SVM*s, *random forest* (RF). For each of these algorithms we also considered the three-way generalization obtained as described in Sect. 3.1 (these algorithms are denoted as TW followed by the acronym of the algorithm as defined previously); in addition, we also considered the three–way decision tree model (in the following denoted as TWDT), described in Sect. 3.2.

4.1 Settings

We compared the algorithms on the following datasets:

- *Iris*: 150 instances, 4 features, 3 classes;
- *Wine*: 178 instance, 13 features, 3 classes;
- *Breast cancer*: 569 instance, 30 features, 2 classes;
- *Digits*: 1797 instances, 64 features, 10 classes;
- *Yeast*: 1484 instances, 8 features, 10 classes;
- *Olivetti faces*: 400 instances, 4096 features, 40 classes;
- *SF12 Mental score* (described in [5]): 462 instances, 10 features, 2 classes;

In order to set the values of τ and ϵ (i.e. the abstention and error costs), we simply selected $\epsilon = 1$ and determined the optimal value of τ using cross-validation. Indeed, for each of the above datasets, we trained the classification algorithms using a 5–fold cross-validation, in order to select the optimal hyper–parameters (which includes τ) of the algorithms (e.g., the tree depth for decision trees). Then, we retrained the algorithms with the best selected hyper–parameters and reported the means and standard deviations of the acc_O accuracy measure (we considered this measure, as motivated in [16], in order to better analyze the trade–off between classification accuracy and abstention). For the three–way classification algorithms, in order to evaluate the trade-off among classification accuracy and coverage (defined as the fraction of objects which are assigned a clear–cut classification), we also measured the *abstention rate*, simply defined as:

$$Abst(A, T) = \sum_{x \in T} \frac{|D(x)|}{|Y|}$$

where A is a three–way classification algorithm, T is a testing set and $D(x) \subseteq Y$, as in Sect. 3.1, is the output labeling of algorithm A on instance x.

In order to more systematically study the trade-off among abstention and classification, for the dataset *Breast cancer* and for algorithms $TWRF$ and $TWSVM$, we also reported the variation with respect to the abstention cost τ of three different metric: accuracy, *true positive rate* (TPR), and *true negative rate* (TNR).

4.2 Results

The results of the experimental comparison are illustrated in Table 1 and, for one specific dataset (Yeast), in Fig. 2.

Table 1. Measured 95% confidence intervals, centered around the mean accuracy, for the considered datasets and algorithms.

Algorithm	Iris	Wine	Breast	Digits	Yeast	Faces	SF12
KNN	0.98 ± 0.03	0.75 ± 0.13	0.93 ± 0.04	0.98 ± 0.03	0.57 ± 0.03	0.90 ± 0.16	0.82 ± 0.02
TWKNN	1.00 ± 0.00	0.99 ± 0.02	0.99 ± 0.01	0.90 ± 0.00	0.67 ± 0.02	0.89 ± 0.01	0.82 ± 0.01
LR	0.95 ± 0.06	0.95 ± 0.05	0.95 ± 0.02	0.93 ± 0.04	0.53 ± 0.03	0.96 ± 0.03	0.73 ± 0.13
TWLR	0.96 ± 0.01	0.98 ± 0.02	0.98 ± 0.01	0.96 ± 0.02	0.78 ± 0.01	0.98 ± 0.01	0.77 ± 0.01
LDA	0.98 ± 0.04	0.98 ± 0.03	0.96 ± 0.03	0.92 ± 0.03	0.59 ± 0.01	0.98 ± 0.01	0.83 ± 0.12
TWLDA	0.98 ± 0.04	0.99 ± 0.00	0.97 ± 0.01	0.94 ± 0.02	0.72 ± 0.07	0.99 ± 0.01	0.83 ± 0.12
NB	0.95 ± 0.04	0.96 ± 0.03	0.94 ± 0.03	0.81 ± 0.06	0.15 ± 0.02	0.82 ± 0.03	0.82 ± 0.07
TWNB	0.97 ± 0.03	0.98 ± 0.03	0.95 ± 0.03	0.83 ± 0.05	0.16 ± 0.02	0.84 ± 0.02	0.86 ± 0.05
SVM	0.98 ± 0.03	0.73 ± 0.09	0.94 ± 0.02	0.97 ± 0.02	0.52 ± 0.03	0.79 ± 0.05	0.74 ± 0.04
TWSVM	0.98 ± 0.01	0.90 ± 0.01	0.96 ± 0.01	1.00 ± 0.00	0.81 ± 0.02	0.87 ± 0.04	0.83 ± 0.06
RF	0.97 ± 0.03	0.98 ± 0.03	0.96 ± 0.02	0.94 ± 0.02	0.58 ± 0.03	0.93 ± 0.02	0.83 ± 0.05
TWRF	0.98 ± 0.01	0.99 ± 0.01	0.99 ± 0.01	0.99 ± 0.01	0.80 ± 0.02	0.98 ± 0.01	0.85 ± 0.03
TWDT	0.97 ± 0.03	0.89 ± 0.07	0.94 ± 0.03	0.83 ± 0.04	0.63 ± 0.02	0.61 ± 0.15	0.84 ± 0.06

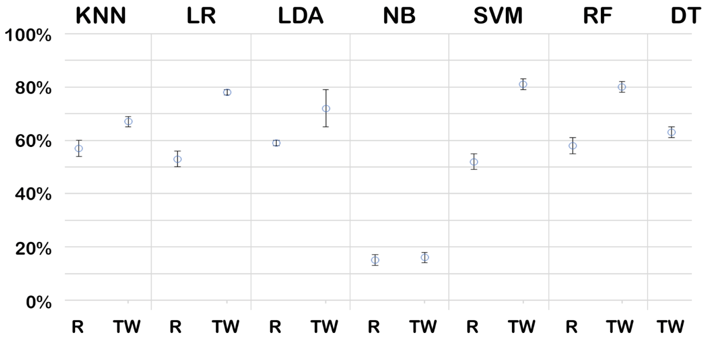

Fig. 2. Measured values of accuracy (and their 95%CIs) for the algorithms under test (regular version, R, on the left) and their three–way version (TW, on the right), on the dataset *Yeast*. Comparing the confidence intervals visually, it is clear that significant differences are observed for 5 model families (namely, KNN, LR, LDA, SVM and RF).

In Table 2, we reported the average ranks of the algorithms (i.e. for each dataset we sorted the algorithms in order of decreasing average accuracy, then we computed the average rank across the datasets).

Table 2. Average ranks of the top 10 performing algorithms.

Alg.	TWRF	TWLDA	TWLR	TWSVM	LDA	RF	TWKNN	KNN	LR	TWNB
Rank	1.75	2.96	3.18	3.57	4.32	4.64	4.64	5.86	6.14	6.78

As can be easily observed from Table 2, in every case the adoption of the possibility of abstention decreases the average rank of the respective algorithm (thus, the algorithm increases its performance). This effect can be explained by noting that the possibility of abstention gives the algorithm the ability to not express a clear–cut decision in those instances which are placed near the decision boundary (i.e. the instances whose class assignment is most uncertain) but, instead, report a list of possible classifications (which, with high confidence, includes the real label). In order to assess if the improvements given by the possibility of abstention were statistically significant, we performed a pair–wise Friedman test [12] for each pair of three–way/classical algorithm, with Li's correction for multiple hypothesis testing [25]: one of the three–way algorithms (TWSVM) was found to be significantly better than the respective classical with a p-value = 0.02, for two others (TWRF, TWLR) there was weak evidence of improvement, albeit with a lower p-value = 0.08 (all other algorithm pairs reported a p–value > 0.1), when considering the standard confidence level of $CL = 95\%$ only the first difference is statistically significant.

In order to investigate the trade–off between classification accuracy and abstention, as mentioned in Sect. 4.1, we measured the abstention rate of the three–way algorithms, as shown in Table 3. It could be easily observed that,

Table 3. Measured abstention rates for the considered datasets and three–way algorithms.

Algorithm	Iris	Wine	Breast	Digits	Yeast	Faces	SF12
TWDT	0.05	0.00	0.08	0.00	0.24	0.08	0.49
TWKNN	0.13	0.58	0.20	0.95	0.11	0.04	0.00
TWLR	0.01	0.19	0.14	0.19	0.27	0.02	0.60
TWLDA	0.00	0.05	0.08	0.30	0.16	0.03	0.00
TWNB	0.07	0.03	0.02	0.05	0.02	0.02	0.34
TWSVM	0.16	0.42	0.05	0.04	0.17	0.11	0.29
TWRF	0.05	0.15	0.13	0.08	0.17	0.05	0.31

in general, the abstention rate is greater than the corresponding increase of accuracy. This effect likely emerges because some of the instances that were classified *correctly* by a classical algorithm, were so *only by chance* (i.e., they were assigned to the correct class label, but with a low confidence level) and, thus, the corresponding three–way algorithm makes this phenomenon apparent (this is particularly evident for the TWKNN, which registered the highest value of abstention rate). An interesting observation is that in the Yeast dataset, the three–way algorithms performed significantly better than the classical ones, with only a moderate increase in abstention rates. It could also be observed that the best performing algorithm (TWRF) was consistently better than the other algorithms in every dataset, although no statistically significant difference (at $CL = 95\%$) could be found with the second ranking algorithm (i.e., TWLDA, $p - value = 0.28$).

Finally, as mentioned in Sect. 4.1, we analyzed the variation of different metrics, that is accuracy, true positive rate (TPR), true negative rate (TNR) and abstention rate, with respect to varying τ on two algorithms: the results are shown in Figs. 3 and 4.

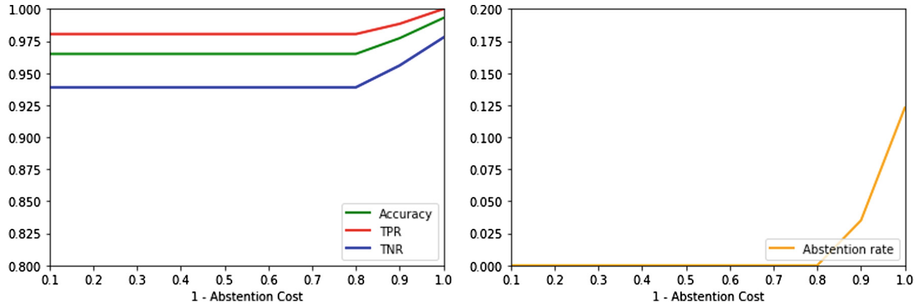

Fig. 3. Variation, w.r.t. abstention cost τ, of different metric for the TWRF algorithm: accuracy, tpr, tnr (left); abstention rate (right).

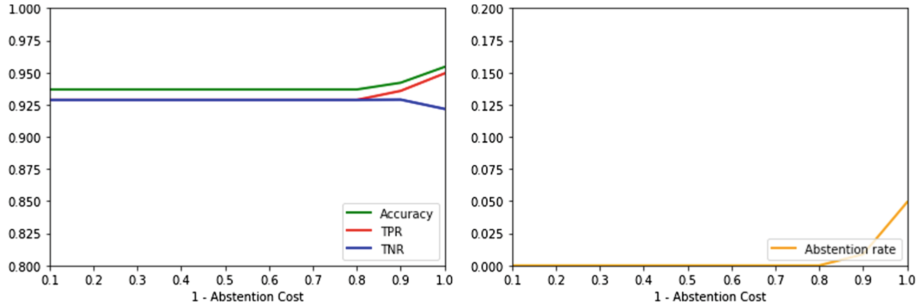

Fig. 4. Variation, w.r.t. abstention cost τ, of different metric for the TWSVM algorithm: accuracy, tpr, tnr (left); abstention rate (right).

As can be easily observed, both the accuracy and the abstention rate increase monotonically with decreasing τ (for both algorithms); furthermore there is a variation in the observed measures only for values $\tau \leq 0.2$ and, even at $\tau = 1$, the observed abstention rates were small; this can be explained, as noted by Theorem 1, as the algorithms assigned great confidence to their predictions.

A final point to note is that the TNR for algorithm TWSVM, shown in Fig. 4, decreases with decreasing abstention cost: this could be related to a deficit in the training dataset, which highlights a possible difficulty in detecting true negative instances.

5 Conclusion

In this work we presented a comprehensive framework to address three–way classification, both in the binary and the multi–class case, by providing a general approach to convert probabilistic classifiers into three–way algorithms. To this aim, we also focused on two techniques to directly embed the possibility of abstention given by this classification approach into three popular learning models. Consequently, in order to evaluate the proposed classification framework, we performed an empirical evaluation comparing a set of traditional learning algorithms with the respective three–way generalizations, on a variety of datasets.

The obtained results showed that, in every case, the possibility to abstain on *difficult instances*, given by three–way classification yields an increase, sometimes significant, in performance and, perhaps more importantly, the possibility to identify the instances that are considered *ambiguous* by the classification algorithms.

This last aspect, in our view, is especially important because it could be used in a *human in the loop* setting, to point out to the human decision–maker which instances might require the acquisition of further or more precise information and require special attention: that is, despite the *uncertainty* intrinsic to these three–way predictions, these could nevertheless be useful to the human decision maker as a way to raise awareness of the weak points and ambiguities affecting the available data.

Given the promising results that we obtained, we plan to continue this line of research considering the following issues and open problems:

- in this paper, we introduced both a general approach to build three–way classifiers and also two more techniques that may be applied to specific learning algorithms. Although we analyzed one such technique (learning of three–way decision trees), we plan to study if directly implementing three–way classification in ensemble tree–based algorithms (e.g. random forests) and convex learning algorithms could be more advantageous than the general post–hoc strategy evaluated in this work;
- in this work, we primarily focused on ambiguity *in the output*, that is, how ambiguity could be managed by allowing three–way, instead of crisp, classifications. However, ambiguity is a multi–faceted problem that could arise also in the input: both in the target attributes (e.g. abstentions are already present in the given gold standard) and the predictor ones (which could present missing or partial values). While we performed some initial works relating to these issues [5,6], we plan to expand this line of research, especially in regard to ambiguity in predictor attributes, in order to build a comprehensive framework for managing ambiguity in machine learning.

References

1. Bartlett, P.L., Wegkamp, M.H.: Classification with a reject option using a hinge loss. J. Mach. Learn. Res. **9**, 1823–1840 (2008)
2. Bello, R., Falcon, R.: Rough sets in machine learning: a review. In: Wang, G., Skowron, A., Yao, Y., Ślęzak, D., Polkowski, L. (eds.) Thriving Rough Sets, pp. 87–118. Springer International Publishing, Cham (2017)
3. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001)
4. Cabitza, F., Ciucci, D., Rasoini, R.: A giant with feet of clay: on the validity of the data that feed machine learning in medicine. In: Cabitza, F., Batini, C., Magni, M. (eds.) Organizing for the Digital World. LNISO, vol. 28, pp. 121–136. Springer, Cham (2019). https://doi.org/10.1007/978-3-319-90503-7_10
5. Campagner, A., Cabitza, F., Ciucci, D.: Exploring medical data classification with three-way decision tree. In: Proceedings of the 12th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2019) - Volume 5: HEALTHINF. pp. 147–158. SCITEPRESS (2019)
6. Campagner, A., Ciucci, D.: Three-way and semi-supervised decision tree learning based on orthopartitions. In: Medina, J., Ojeda-Aciego, M., Verdegay, J.L., Pelta, D.A., Cabrera, I.P., Bouchon-Meunier, B., Yager, R.R. (eds.) IPMU 2018. CCIS, vol. 854, pp. 748–759. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-91476-3_61
7. Campagner, A., Ciucci, D.: Orthopartitions and soft clustering. Knowl. Based Syst. (Submitted)
8. Chow, C.: On optimum recognition error and reject tradeoff. IEEE Trans. Inform. Theory **16**, 41–46 (1970)
9. Ciucci, D.: Orthopairs: a simple and widely used way to model uncertainty. Fundamenta Informaticae **108**, 287–304 (2011)

10. Ciucci, D.: Orthopairs and granular computing. Granular Comput. **1**, 159–170 (2016)
11. Cortes, C., Vapnik, V.: Support-vector networks. Mach. Learn. **20**(3), 273–297 (1995)
12. Daniel, W.W.: Applied Nonparametric Statistics. Duxbury Thomson Learning (1990)
13. Deo, R.: Machine learning in medicine. Circulation **132** (2015)
14. Ellerman, D.: An introduction to logical entropy and its relation to Shannon entropy. Int. J. Semant. Comput. **7**(2), 121–145 (2013)
15. Feldman, K., Faust, L., Wu, X., Huang, C., Chawla, N.V.: Beyond volume: the impact of complex healthcare data on the machine learning pipeline. CoRR abs/1706.01513 (2017)
16. Ferri, C., Hernández-Orallo, J.: Cautious classifiers. In: ROC Analysis in Artificial Intelligence, 1st International Workshop, ROCAI-2004, pp. 27–36 (2004)
17. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press, Cambridge (2016)
18. Hajian, S., Bonchi, F., Castillo, C.: Algorithmic bias: from discrimination discovery to fairness-aware data mining. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 2125–2126, August 2016
19. Han, P.K., Klein, W.M., Arora, N.K.: Varieties of uncertainty in health care: a conceptual taxonomy. Med. Decis. Making **31**(6), 828–838 (2011)
20. Hechtlinger, Y., Póczos, B., Wasserman, L.A.: Cautious deep learning. arXiv/CoRR abs/1805.09460 (2018)
21. Hüllermeier, E.: Fuzzy sets in machine learning and data mining. Appl. Soft Comput. **11**(2), 1493–1505 (2011)
22. Hüllermeier, E.: Does machine learning need fuzzy logic? Fuzzy Sets Syst. **281**, 292–299 (2015). Special Issue Celebrating the 50th Anniversary of Fuzzy Sets
23. Koller, D., Friedman, N.: Probabilistic Graphical Models: Principles and Techniques - Adaptive Computation and Machine Learning. The MIT Press, Cambridge (2009)
24. Kooi, T., et al.: Large scale deep learning for computer aided detection of mammographic lesions. Med. Image Anal. **35**, 303–312 (2017)
25. Li, J.D.: A two-step rejection procedure for testing multiple hypotheses. J. Stat. Plann. Infer. **138**(6), 1521–1527 (2008)
26. Obermeyer, Z., Emanuel, E.J.: Predicting the future - big data, machine learning, and clinical medicine. N. Engl. J. Med. **375**(13), 1216–1219 (2016)
27. Pawlak, Z.: Rough sets. Int. J. Comput. Inform. Sci. **11**(5), 341–356 (1982)
28. Shafer, G.: A Mathematical Theory of Evidence. Princeton University Press, Princeton (1976)
29. Smets, P., Kennes, R.: The transferable belief model. Artif. Intell. **66**(2), 191–234 (1994)
30. Svensson, C., Hübler, R., Figge, M.: Automated classification of circulating tumor cells and the impact of interobserver variability on classifier training and performance. J. Immunol. Res. **2015**, 1–9 (2015)
31. Yao, Y.: An outline of a theory of three-way decisions. In: Yao, J.T., Yang, Y., Słowiński, R., Greco, S., Li, H., Mitra, S., Polkowski, L. (eds.) RSCTC 2012. LNCS (LNAI), vol. 7413, pp. 1–17. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-32115-3_1
32. Zadeh, L.: Fuzzy sets. Inf. Control **8**(3), 338–353 (1965)