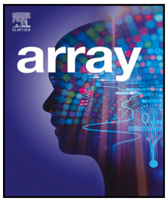
Array 19 (2023) 100310

Contents lists available at ScienceDirect

Array

journal homepage: www.elsevier.com/locate/array



The study of the hyper-parameter modelling the decision rule of the cautious classifiers based on the F_β measure

Abdelhak Imoussaten

EuroMov Digital Health in Motion, Univ Montpellier, IMT Mines Ales, Ales, France

ARTICLE INFO

Keywords:
Cautious classification
Set-valued classification
Belief functions
Supervised machine learning

ABSTRACT

In some sensitive domains where data imperfections are present, standard classification techniques reach their limits. To avoid misclassifications that have serious consequences, recent works propose cautious classification algorithms to handle this problem. Despite of the presence of uncertainty and/or imprecision, a point prediction classifier is forced to bet on a single class. While a cautious classifier proposes the appropriate subset of candidate classes that can be assigned to the sample in the presence of imperfect information. On the other hand, cautiousness should not be at the expense of precision and a trade-off has to be made between these two criteria. Among the existing cautious classifiers, two classifiers propose to manage this trade-off in the decision step by the mean of a parametrized objective function. The first one is the non-deterministic classifier (ndc) proposed within the framework of probability theory and the second one is "evidential classifier based on imprecise relabelling" (eclair) proposed within the framework of belief functions. The theoretical aim of the mentioned hyper-parameters is to control the size of predictions for both classifiers. This paper proposes to study this hyper-parameter in order to select the "best" value in a classification task. First the utility for each candidate subset is studied related to the values of the hyper-parameter and some thresholds are proposed to control the size of the predictions. Then two illustrations are proposed where a method to choose this hyper-parameters based on the calibration data is proposed. The first illustration concerns randomly generated data and the second one concerns the images data of fashion mnist. These illustrations show how to control the size of the predictions and give a comparison between the performances of the two classifiers for a tuning based on our proposition and the one based on grid search method.

1. Introduction

In some sensitive applications misclassification can have serious consequences. This is the case in applications having impacts either on people's health or on the environment [1], e.g., in medical diagnosis applications when a classifier is involved to detect early-stage cancer. In such applications, cautious decisions are necessary when imperfect data are present. This leads some recent works to focus on cautious classification. Among the existing cautious classifiers, we focus on those providing a subset of candidate class labels to a new sample to classify. We refer to those classifiers as set-valued classifiers. We can cite among these classifiers the *Naive Credal Classifier* (*ncc*) [2,3], the strong dominance based classifier [4], the *non-deterministic* classifier (ndc) [5], the credal decision trees (CDT) [6], the imprecise credal decision trees (ICDT) [7], the classifiers based on generalized criteria such as Hurwicz (GHC), OWA (GOWAC) [8], eclair classifier [9,10], etc. One can find in [8–10] other works concerning cautious prediction including works about conformal prediction [11–14] that are discussed in this paper. But cautiousness should not be at the expense of precision and a trade-off has to be made between these two criteria. On one hand, a classifier that predicts always the whole set of the candidate classes is cautious but its predictions are uninformative. On the other hand, a classifier that predicts always a single class is precise when the predictions are good but it is not cautious. Some set-valued classifiers can control this trade-off as the *ndc* and *eclair* classifiers. Indeed, the utility function implemented in the decision step of both classifiers has a hyper-parameter β that is used to control the trade-off between precision and cautiousness. This hyper parameter is considered as a user-modifiable parameter for the application of those classifiers and its theoretical aim is to control the size of the predicted subset of classes. The choice of β depends on the level of cautiousness required for the application in which the classifier is going to be used. This paper proposes to study this parameter in the case of the two classifiers and aims to propose suggestions for the choice of the parameter value in the case of classification task. In the first experiment results, we show, on simulated data, the impact of the selected hyper-parameter value on the prediction of the two classifiers when faced to strange samples, i.e., to

E-mail address: abdelhak.imoussaten@mines-ales.fr.

https://doi.org/10.1016/j.array.2023.100310
Received 1 April 2023; Received in revised form 5 July 2023; Accepted 17 July 2023
Available online 20 July 2023

which the standard classifiers failed to predict the true class labels. In the second experiment, concerning the images data of fashion mnist, a comparison between the performances of the two classifiers for a tuning based on our proposition and the one based on grid search method is presented. The paper is organized as follows. In the second section, the reminders about the decision step in the classifiers *eclair* and *ndc* and the measures of set-valued classification performances are given. The third section presents the studies of the expected utility functions introduced in the decision step of the two classifiers. Finally, the fourth section presents the experimental results.

2. Reminders and notations

The set-valued classifiers *eclair* and *ndc* are based on the results of the standard point prediction classifiers to provide respectively the posterior mass function and the posterior probability function for a sample $\boldsymbol{x}$ to classify among a set of classes $\Theta = \{\theta_1, \dots, \theta_n\}$. We focus in this paper on the decision step of those two classifiers that involves a utility function that is the F_β score. In this section we give some reminders about the F_β score and it exploitation in the case of imprecise predictions by the two classifiers. The evaluation of imprecise predictions is performed using five measures from the state of the art that are presented in the end of this section. To simplify notations, we adopt the following notations for the singleton subsets and subsets of two elements, in the rest of the paper: $\theta_i := \{\theta_i\}$, $\theta_{ij} := \{\theta_i, \theta_j\}$.

2.1. F_β score

The F_β score used in the decision step of *eclair* and *ndc* to predict a subset of candidate classes is an adaptation of the F_β score from information retrieval field and supervised classification methods to set-valued classification. In the context of binary point prediction for classification, the F_β score is defined as:

$$F_\beta(predictions, truth) = \frac{(1+\beta^2)\ \mathrm{recall} \cdot \mathrm{precision}}{(\beta^2 \cdot \mathrm{precision}) + \mathrm{recall}}, \tag{1}$$

where the quantity *precision* defined as:

$$precision = \frac{nb\ of\ true\ predicted\ as\ true}{nb\ of\ true\ predicted\ as\ true + nb\ of\ false\ predicted\ as\ true},$$

and the quantity *recall* defined as:

$$recall = \frac{nb\ of\ true\ predicted\ as\ true}{nb\ of\ true\ predicted\ as\ true + nb\ of\ true\ predicted\ as\ false},$$

are two known performance measures in information retrieval and machine learning. This general expression of F_β score in Eq. (1) is parametrized by β that is chosen such that *recall* is considered β times as important as *precision*.

2.2. The decision step in ndc classifier

The principle of *ndc* is very simple. Let us consider a sample $\boldsymbol{x}$ and a trained point prediction model, denoted δ, that can provide a posterior probability for the classification of $\boldsymbol{x}$. The *ndc* classifier consists in a decision rule r_{ndc} that is applied to determine the set-valued prediction for $\boldsymbol{x}$. Note that, precise predictions are given as singleton subsets. The predicted subset of classes, using the rule r_{ndc}, is the one maximizing the expected utility where the utility associated to each subset of classes is defined using the F_β measure. More precisely, let us consider a set of n class labels $\Theta = \{\theta_1, \dots, \theta_n\}$, the utility of each subset of candidate classes $A \subseteq \Theta$ as the good prediction for $\boldsymbol{x}$, having the true class $\theta_{\boldsymbol{x}}$, is evaluated using the F_β measure as follows:

$$F_\beta(A, \theta_{\boldsymbol{x}}) = \frac{(1+\beta^2) \cdot |\{\theta_{\boldsymbol{x}}\} \cap A|}{\beta^2 + |A|}, \tag{2}$$

where β is a positive real number and $|X|$, for $X \subseteq \Theta$, denotes the number of elements in X. The quantity $F_\beta(A, \theta_{\boldsymbol{x}})$ is interpreted as the utility obtained when predicting the subset of class labels A when the true class label is $\theta_{\boldsymbol{x}}$. Eq. (2) is analogue to the one in (1) where the quantities *precision* and *recall* are redefined as $\mathrm{precision}(A) = \frac{|\{\theta_{\boldsymbol{x}}\} \cap A|}{nb\, of\ classes\, in A}$ and $\mathrm{recall}(A) = |\{\theta_{\boldsymbol{x}}\} \cap A|$. Note that when the value of β is close to 0, $F_\beta(A, \theta_{\boldsymbol{x}})$ becomes close to $\mathrm{precision}(A)$ thus the size of A is disadvantageous, i.e. the larger size of A the smaller the utility. On the other hand, when β is high, $F_\beta(A, \theta_{\boldsymbol{x}})$ becomes close to $\mathrm{recall}(A)$ and in this case the size of A is advantageous. Let us suppose that a posterior probability distribution $p_\delta(.|\boldsymbol{x})$ is provided by a point prediction method δ for the sample $\boldsymbol{x}$, then the non-deterministic classifier *ndc* predicts for $\boldsymbol{x}$ the subset of candidate classes that maximize the expected utility function $\mathbb{E}(F_\beta(A, .)|\boldsymbol{x})$, i.e. :

$$\mathbb{E}(F_\beta(A, .)|\boldsymbol{x}) = \sum_{i=1}^{n} F_\beta(A, \theta_i) \cdot p_\delta(\theta_i|\boldsymbol{x}). \tag{3}$$

Finally, the predicted subset $r_{ndc}(\boldsymbol{x})$ for $\boldsymbol{x}$ using the classifier *ndc* is given as:

$$r_{ndc}(\boldsymbol{x}) = arg \max_{A \in 2^\Theta \setminus \emptyset} \mathbb{E}(F_\beta(A, .)|\boldsymbol{x}). \tag{4}$$

2.3. The decision step in eclair classifier

The decision step with *eclair* consists in providing for a sample $\boldsymbol{x}$ a subset of classes as prediction, by considering as input the posterior mass function $m(.|\boldsymbol{x})$ and a hyper-parameter β. The predicted subset of classes is the one maximizing the expected utility where the utility associated to each subset of classes is defined using a generalization of Eq. (2) [9,10]. The main change regarding Eq. (2) is to consider the general case where the available information about the true class of $\boldsymbol{x}$ can be partially known, i.e., a subset $B_{\boldsymbol{x}}$ of Θ. It is the case, for example, when data are coarse [15,16]. The new utility function is then defined for two subsets A and $B_{\boldsymbol{x}}$ of Θ as follows:

$$F_\beta(A, B_{\boldsymbol{x}}) = \frac{(1+\beta^2) \cdot |A \cap B_{\boldsymbol{x}}|}{\beta^2 \cdot |B_{\boldsymbol{x}}| + |A|} \tag{5}$$

The quantity $F_\beta(A, B_{\boldsymbol{x}})$ is interpreted as the utility obtained when predicting the subset of class labels A for the sample $\boldsymbol{x}$ when its true class label is partially known and represented by a subset of classes $B_{\boldsymbol{x}}$. In this case, the *precision* and *recall* quantities become:

$$\mathrm{precision}(A) = \frac{|A \cap B_{\boldsymbol{x}}|}{nb\, of\ classes\, inA},$$

and

$$\mathrm{recall}(A) = \frac{|A \cap B_{\boldsymbol{x}}|}{nb\, of\ classes\ in\ B_{\boldsymbol{x}}}.$$

Let us suppose that a posterior mass function $m(.|\boldsymbol{x})$ is known for the sample $\boldsymbol{x}$, the *eclair* classifier predicts for $\boldsymbol{x}$ the subset of candidate classes that maximize the expected utility function $\mathbb{E}(F_\beta(A, .)|\boldsymbol{x})$, i.e :

$$\mathbb{E}(F_\beta(A, .)|\boldsymbol{x}) = \sum_{B \subseteq \Theta} F_\beta(A, B) \cdot m(B|\boldsymbol{x}), \tag{6}$$

where B is the variable representing the true class label of $\boldsymbol{x}$. Finally, the predicted subset $r_{eclair}(\boldsymbol{x})$ for $\boldsymbol{x}$ using the classifier *eclair* is given as:

$$r_{eclair}(\boldsymbol{x}) = arg \max_{A \in 2^\Theta \setminus \emptyset} \mathbb{E}(F_\beta(A, .)|\boldsymbol{x}). \tag{7}$$

2.4. Evaluation measures for the set-valued classifiers

When evaluating a set-valued classifier one ensures that the predicted subset of classes (1) includes the "true" class and (2) it is as small as possible depending on the sample data imperfection. Several works have studied this problem and provide some measures to check the two conditions (1) and (2) [2,7,17]. Between the least drastic one that is *imprecise accuracy* which checks if the prediction contains the true class label of the sample and the most drastic one that is *classical accuracy* which checks if the prediction is equal to the true class label of the sample, one can find intermediate measure as *Discounted accuracy* [18] that seems to be an interesting measure as it takes into account the size of the predicted subset. But in order to increase the cautiousness reward to the degree to which the decision maker prefers to fix it depending on his application and the quality of the information obtained for the samples, a family of measure are constructed from *Discounted accuracy* measure that are represented by a function g taking its values in $[0,1]$ and guaranteeing $g(z) \geq z$, i.e., the reward with g is at least the same as the one given by the *discounted accuracy*, $g(0) = 0$ and $g(1) = 1$ (see [19] for more details). Let us consider a samples $\boldsymbol{x}$ having the true class $\theta_{\boldsymbol{x}}$ and a set-valued classifier δ_{ic}. The five following measures are proposed to evaluate the performance of δ_{ic} regarding its output for $\boldsymbol{x}$:

- the *classical accuracy* denoted *acc*:

  $$acc(\delta_{ic}, \theta_{\boldsymbol{x}}) = \mathbb{1}_{\{\theta_{\boldsymbol{x}}\}}(\delta_{ic}(\boldsymbol{x})).$$

- the *imprecise accuracy* denoted *impr. acc*:

  $$impr.acc(\delta_{ic}, \theta_{\boldsymbol{x}}) = \mathbb{1}_{\delta_{ic}(\boldsymbol{x})}(\theta_{\boldsymbol{x}}).$$

- the *discounted accuracy* (discAcc) corresponds to the function $g(z) = z$ [18]:

  $$discAcc(\delta_{ic}, \theta_{\boldsymbol{x}}) = \frac{\mathbb{1}_{\delta_{ic}(\boldsymbol{x})}(\theta_{\boldsymbol{x}})}{|\delta_{ic}(\boldsymbol{x})|},$$

  where $|A|$ denotes the size of the subset A. This measure is also denoted u_{50}.
- The u_{65} measure that corresponds to the function $g(z) = -0.6 \cdot z^2 + 1.6 \cdot z$ [19]:

  $$u_{65}(\delta_{ic}, dst) = -0.6 \cdot [discAcc(\delta_{ic}, \theta_{\boldsymbol{x}})]^2 + 1.6 \cdot discAcc(\delta_{ic}, \theta_{\boldsymbol{x}}).$$

- The u_{80} measure that corresponds to the function $g(z) = -1.2 \cdot z^2 + 2.2 \cdot z$ [19]:

  $$u_{80}(\delta_{ic}, \theta_{\boldsymbol{x}}) = -1.2 \cdot [discAcc(\delta_{ic}, \theta_{\boldsymbol{x}})]^2 + 2.2 \cdot discAcc(\delta_{ic}, \theta_{\boldsymbol{x}}).$$

Note that u_{65} is the average measure of u_{50} and u_{80} and it is the better suited one to quantify the compromise between precision and cautiousness.

3. The expected utilities related to β

3.1. *The case of ndc classifier*

Let us consider that the posterior probability distribution of a sample $\boldsymbol{x}$ is known. We denote this distribution by $p(.|\boldsymbol{x}) : \Theta \to [0,1]$. We consider the parameter β as a variable and we express the expected utility function in Section 2.2 for a $\beta \in [0,+\infty[$, $A \subseteq \Theta$ and $p(.|\boldsymbol{x})$ as:

$$u(\beta, A) = \mathbb{E}(F_\beta(A,.)|\boldsymbol{x}) = \sum_{i=1}^{n} F_\beta(A, \theta_i) \cdot p(\theta_i|\boldsymbol{x}). \tag{8}$$

In addition, let us consider the situation where the class θ_k is the most likely class of $\boldsymbol{x}$ and some times the class θ_k is confused with the class $\theta_{k'}$, $k \neq k'$ due to data imperfection. The Propositions 3.1 and 3.2 give some results concerning the predicted subset of classes for $\boldsymbol{x}$ among the three options θ_k, $\theta_{kk'}$ and Θ.

Proposition 3.1. *Let suppose that $p(\theta_k|\boldsymbol{x}) > p(\theta|\boldsymbol{x})$, $\forall \theta \in \Theta \setminus \theta_k$ and $\theta_{k'} = arg\max_{\theta \in \Theta \setminus \theta_k} p(\theta|\boldsymbol{x})$.*

If $p(\theta_{k'}|\boldsymbol{x}) > 0$ then it exists $\beta_1 \geq 0$ such that:

$$\begin{cases} u(\beta, \theta_{kk'}) \leq u(\beta, \theta_k) & \text{if } \beta \leq \beta_1 \\ u(\beta, \theta_{kk'}) > u(\beta, \theta_k) & \text{if } \beta > \beta_1. \end{cases} \tag{9}$$

Elsewhere $u(\beta, \Theta) < u(\beta, \theta_{kk'})$, $\forall \beta \geq 0$.

Proof. We have for all $\beta \geq 0$,

$$u(\beta, \theta_k) = p(\theta_k|\boldsymbol{x}).$$

Table 1
The posterior probability distributions.

x	θ_1	θ_2	θ_3
x_1	0.333	0.333	0.333
x_2	1	0	0
x_3	0.5	0.5	0
x_4	0.5	0.4	0.1

and

$$u(\beta, \theta_{kk'}) = \frac{1+\beta^2}{2+\beta^2} \cdot [p(\theta_k|\boldsymbol{x}) + p(\theta_{k'}|\boldsymbol{x})] = \frac{1+\beta^2}{2+\beta^2} \cdot \mathbb{P}(\theta_{kk'}|\boldsymbol{x}),$$

where for all subset A of Θ, $\mathbb{P}(A|\boldsymbol{x}) = \sum_{\theta \in A} p(\theta|\boldsymbol{x})$. On one hand, the function $u(., \theta_{kk'})$ increases related to β. Thus $u(\beta, \theta_{kk'}) \geq u(0, \theta_{kk'}) = \frac{1}{2}\mathbb{P}(\theta_{kk'}|\boldsymbol{x})$, for all $\beta \geq 0$. On the other hand, $p(\theta_k|\boldsymbol{x}) > p(\theta_{k'}|\boldsymbol{x})$ then $p(\theta_k|\boldsymbol{x}) > \frac{1}{2}\mathbb{P}(\theta_{kk'}|\boldsymbol{x})$. So, $u(., \theta_{kk'})$ intersects $u(., \theta_k)$ at $\beta_1 \geq 0$ such that:

$$\frac{1+\beta_1^2}{2+\beta_1^2} \cdot \mathbb{P}(\theta_{kk'}|\boldsymbol{x}) = p(\theta_k|\boldsymbol{x}).$$

It comes:

$$\beta_1 = \sqrt{\frac{p(\theta_k|\boldsymbol{x}) - p(\theta_{k'}|\boldsymbol{x})}{p(\theta_{k'}|\boldsymbol{x})}}. \quad \square \tag{10}$$

Note that the same reasoning for the comparison between the utilities of θ_k and $\theta_{kk'}$ in Proposition 3.1 can be generalized for the comparisons between the utilities of θ_k and all the subsets $A \subset \Theta$ containing θ_k where

$$u(\beta, A) = \frac{1+\beta^2}{|A|+\beta^2} \cdot \mathbb{P}(A|\boldsymbol{x}),$$

in this case, β_1 becomes:

$$\beta_1 = \sqrt{\frac{|A| \cdot p(\theta_k|\boldsymbol{x}) - \mathbb{P}(A|\boldsymbol{x})}{\mathbb{P}(A|\boldsymbol{x}) - p(\theta_k|\boldsymbol{x})}}. \tag{11}$$

Proposition 3.2. *Let suppose that $p(\theta_k|\boldsymbol{x}) > p(\theta|\boldsymbol{x})$, $\forall \theta \in \Theta \setminus \theta_k$.*

If $\mathbb{P}(\theta_{kk'}|\boldsymbol{x}) \in [\frac{2}{3}, 1[$ then it exists $\beta_2 > 0$ such that:

$$\begin{cases} u(\beta, \Theta) \leq u(\beta, \theta_{kk'}) & \text{if } \beta \leq \beta_2 \\ u(\beta, \Theta) > u(\beta, \theta_{kk'}) & \text{if } \beta > \beta_2. \end{cases} \tag{12}$$

Proof. We have for all $\beta \geq 0$,

$$u(\beta, \Theta) = \frac{1+\beta^2}{3+\beta^2},$$

and

$$u(\beta, \Theta) - u(\beta, \theta_{kk'}) = \frac{(1+\beta^2) \cdot (2 - 3 \cdot \mathbb{P}(\theta_{kk'}) + (1 - \mathbb{P}(\theta_{kk'}) \cdot \beta^2))}{(3+\beta^2) \cdot (2+\beta^2)}.$$

If $\mathbb{P}(\theta_{kk'}|\boldsymbol{x}) < \frac{2}{3}$, then $u(\beta, \Theta) > u(\beta, \theta_{kk'})$, $\forall \beta \geq 0$. Else, if $\mathbb{P}(\theta_{kk'}|\boldsymbol{x}) = 1$, then $u(\beta, \Theta) = \frac{1+\beta^2}{3+\beta^2} < \frac{1+\beta^2}{2+\beta^2} = u(\beta, \theta_{kk'})$, $\forall \beta \geq 0$. Otherwise, let us consider the following value $\beta^* \geq 0$ such that:

$$\beta^{*2} = \frac{3\ \mathbb{P}(\theta_{kk'}|\boldsymbol{x}) - 2}{1 - \mathbb{P}(\theta_{kk'}|\boldsymbol{x})}, \tag{13}$$

the $\beta_2 = \beta^*$ verify the inequalities of Proposition 3.2. $\square$

Example 3.1. Let us consider the case where $\Theta = \{\theta_1, \theta_2, \theta_3\}$. The posterior probabilities of four samples are given in Table 1 and in Fig. 1. These distributions express several situations of sharing the probability masses between the three classes. For the first sample x_1 the mass is uniformly distributed between the classes; for x_2 the total mass is given to the class θ_1; for x_3 the mass is uniformly distributed between θ_1 and θ_2; and for x_4 the mass distribution is as follows $p(\theta_3|\boldsymbol{x}_4) < p(\theta_1|\boldsymbol{x}_4) < p(\theta_2|\boldsymbol{x}_4)$. As one can see in Fig. 1, for the samples x_1, x_2 and

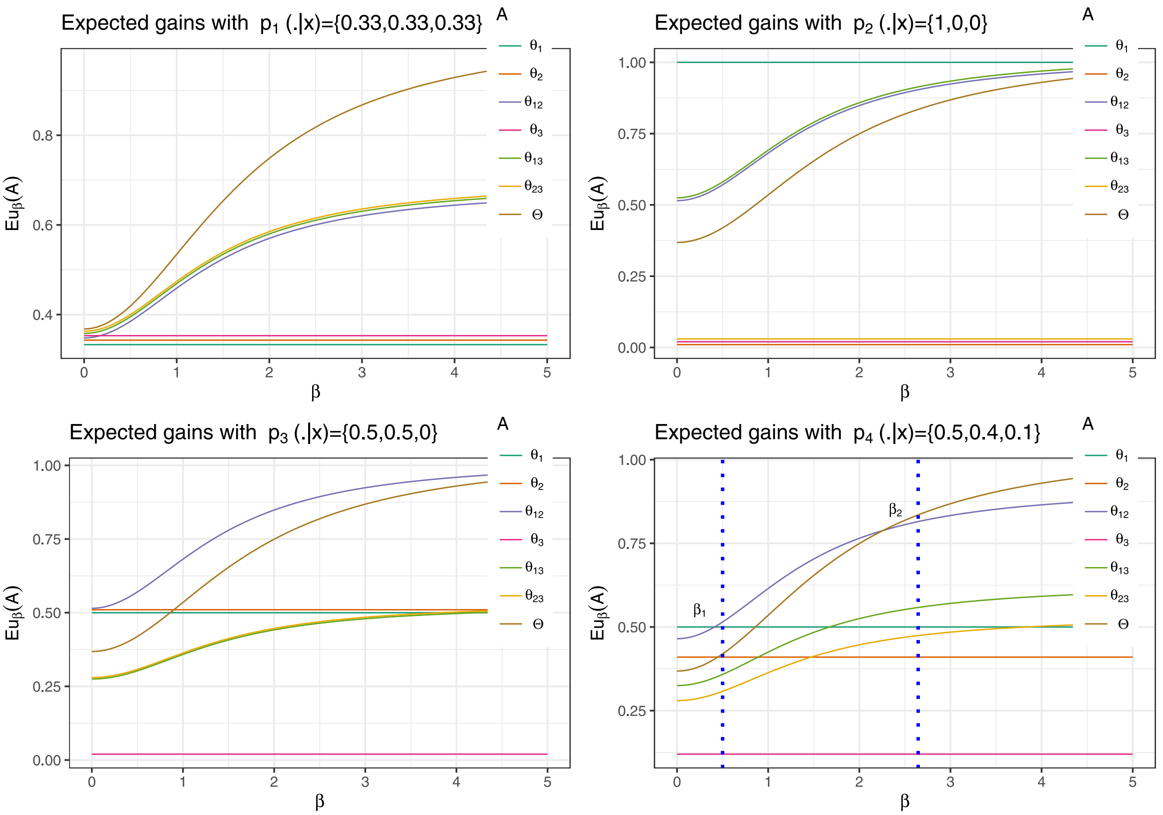


Fig. 1. The expected utility associated to the four posterior probabilities.

x_3, Θ, θ_1, and $\theta_{1,2}$ are respectively the predictions as they maximize the expected utility regardless the value of β. While in the case of x_4, the prediction depends on the value of the parameter β. Indeed, if $\beta < \beta_1 = \sqrt{\frac{p(\theta_1|x_4)-p(\theta_2|x_4)}{p(\theta_2|x_4)}} = 0.5$, i.e., the value of β where the curves of $u(.,\theta_1)$ and $u(.,\theta_{1,2})$ intersect, then θ_1 dominates all the other options. When $\beta_2 > \beta > \beta_1$ ($\beta_2 = \sqrt{\frac{3\,\mathbb{P}(\theta_{1,2}|\boldsymbol{x})-2}{1-\mathbb{P}(\theta_{1,2}|\boldsymbol{x})}} = 2.65$), then $\theta_{1,2}$ dominates all the other options. When $\beta \geq \beta_2$, it is the turn of Θ to dominate the other options.

3.2. *The case of eclair classifier*

In this subsection, we consider that the posterior mass function of a sample $\boldsymbol{x}$ is known. We denote this mass function by $m(.|\boldsymbol{x}) : 2^{\Theta} \to [0,1]$. In this case, the expected utility function used as the criterion to choose the subset of classes to associate to $\boldsymbol{x}$ is the following:

$$u_m(\beta, A) = \mathbb{E}(F_\beta(A,.)|\boldsymbol{x}) = \sum_{B \subseteq \Theta} F_\beta(A,B) \cdot m(B|\boldsymbol{x}), \tag{14}$$

In this section, we treat only the case of two classes. Consequently, the multi-class case can be treated using one-against-one prediction techniques and then infer the final prediction by merging all the one-against-one predictions.

Proposition 3.3. *Let us consider the case where $\Theta = \{\theta_1, \theta_2\}$. If $m(\theta_1|\boldsymbol{x}) > m(\theta_2|\boldsymbol{x})$, then it exists $\beta_3 \geq 0$ such that:*

$$\begin{cases} u_m(\beta,\theta_{12}) \leq u(\beta,\theta_1) & \text{if } \beta \leq \beta_3 \\ u_m(\beta,\theta_{12}) > u(\beta,\theta_1) & \text{if } \beta > \beta_3 \end{cases} \tag{15}$$

Elsewhere, $u_m(\beta,\theta_{12}) \geq u(\beta,\theta_1)$, $\forall \beta \geq 0$.

Proof. In one hand, we have,

$$\frac{du_m(\beta,\theta_1)}{d\beta} = -\frac{2\,\beta}{(1+2\,\beta^2)^2} m(\theta_{12}|\boldsymbol{x})$$

consequently $u_m(.,\theta_1)$ decreases $\forall \beta \geq 0$ with $u_m(0,\theta_1) = m(\theta_1|\boldsymbol{x}) + m(\theta_{12}|\boldsymbol{x})$ and $\lim_{\beta\to+\infty} u_m(\beta,\theta_1) = m(\theta_1|\boldsymbol{x}) + \frac{m(\theta_{12}|\boldsymbol{x})}{2}$. In the other hand, we have,

$$\frac{du_m(\beta,\theta_{12})}{d\beta} = \frac{2\,\beta}{(2+\beta^2)^2}[1 - m(\theta_{12}|\boldsymbol{x})]$$

consequently $u_m(.,\theta_{12})$ increases $\forall \beta \geq 0$ with $u_m(0,\theta_{12}) = \frac{1}{2} + \frac{m(\theta_{12}|\boldsymbol{x})}{2}$ and $\lim_{\beta\to+\infty} u_m(\beta,\theta_{12}) = 1$. Obviously, if $u_m(0,\theta_1) > u_m(0,\theta_{12})$ then $u_m(.,\theta_1)$ and $u_m(.,\theta_{12})$ intersect, elsewhere $u_m(\beta,\theta_{12}) \geq u_m(\beta,\theta_1)$, $\forall \beta \geq 0$. The inequality $u_m(0,\theta_1) > u_m(0,\theta_{12})$ corresponds to $m(\theta_1|\boldsymbol{x}) + m(\theta_{12}|\boldsymbol{x}) > \frac{1}{2} + \frac{m(\theta_{12}|\boldsymbol{x})}{2}$ which is verified when $m(\theta_1|\boldsymbol{x}) > m(\theta_2|\boldsymbol{x})$. Finally, β_3 is the solution of $u_m(\beta,\theta_1) = u_m(\beta,\theta_{12})$ which corresponds to the solution of Eq. (16):

$$m(\theta_1|\boldsymbol{x}) + \frac{1+\beta^2}{1+2\,\beta^2} m(\theta_{12}|\boldsymbol{x}) = \frac{1+\beta^2}{2+\beta^2} + \frac{1}{2+\beta^2} m(\theta_{12}|\boldsymbol{x}). \quad \square \tag{16}$$

Remark 3.1. Note that when m is a Bayesian mass function, i.e., $m(\theta_{12}|\boldsymbol{x}) = 0$, Eq. (16) becomes: $m(\theta_1|\boldsymbol{x}) = \frac{1+\beta^2}{2+\beta^2}$, which is verified for the following value of β_3:

$$\beta_3 = \beta_1 = \sqrt{\frac{m(\theta_1|\boldsymbol{x}) - m(\theta_2|\boldsymbol{x})}{m(\theta_2|\boldsymbol{x})}}.$$

Example 3.2. To illustrate different situations, we consider six mass functions (see Table 2 and Fig. 2). Fig. 2 shows that when $m(\theta_1|x) = m(\theta_2|x)$, e.g. m_1 and m_4, regardless the mass of θ_{12}, the option θ_{12} obtains the maximal gains for all $\beta > 0$. In the other cases, the value of β_3 depends on the mass of θ_{12}, i.e, ignorance. Indeed, the higher the mass of ignorance, the smaller the value of β_3. This means that if the decision-maker desire to make precise predictions for examples like those, he needs to use very small value of β_3 lower than the solution of Eq. (16).

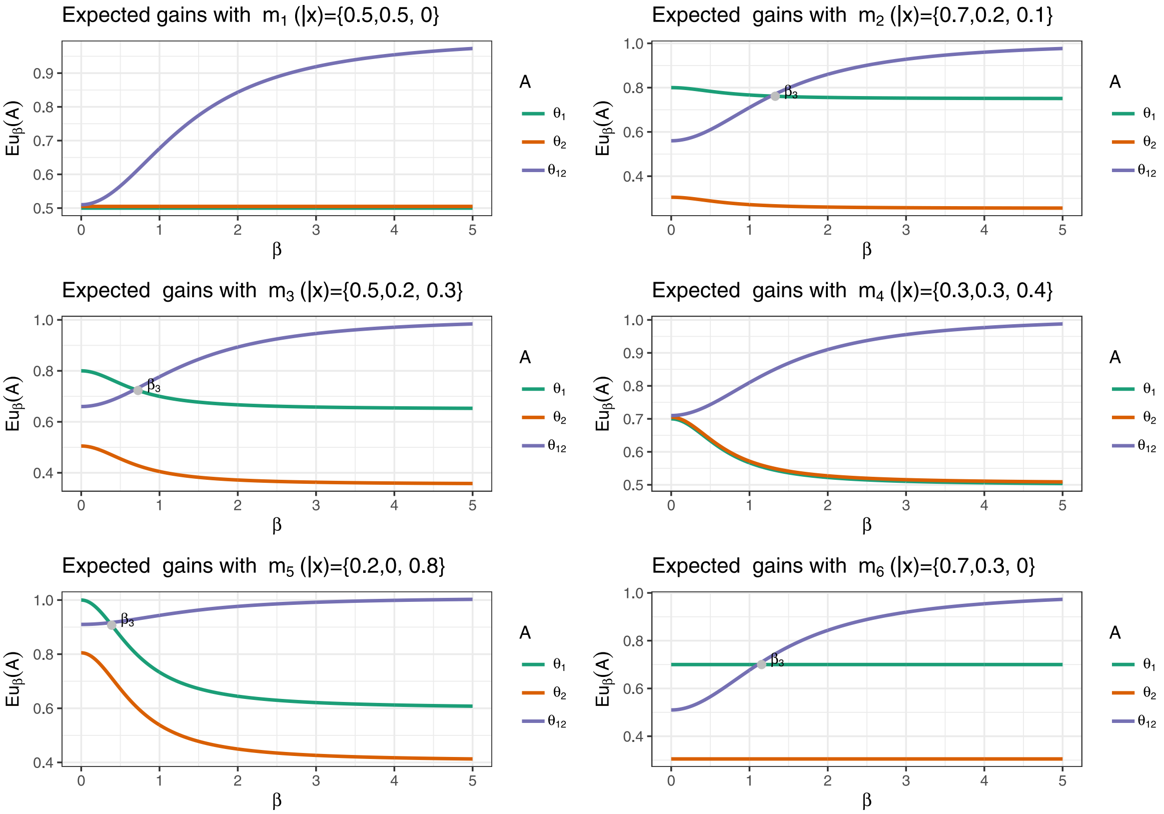


Fig. 2. The utility function for some examples of masses.

Table 2
Mass functions representing several uncertainty situations.

	θ_1	θ_2	$\theta_{1,2}$		θ_1	θ_2	$\theta_{1,2}$
m_1	0.5	0.5	0	m_4	0.3	0.3	0.4
m_2	0.7	0.2	0.1	m_5	0.2	0	0.8
m_3	0.5	0.2	0.3	m_6	0.7	0.3	0

4. Illustration

In this section we present the illustration of the performances of the classifiers *ndc* and *eclair* using generated data in Section 4.1 and using fashion mnist data in Section 4.2. In the two subsections we give the comparisons of the classifiers when the hyper-parameter is tuned based on grid search method or based on the proposition of this paper regarding the set-valued classification metrics and in Section 4.2 we show how to control the number of the predictions using our propositions.

4.1. Illustration using simulated data

In this first illustration, we consider a simulated data for three class labels a, b, and c. For each class label 500 training samples of a bivariate Gaussian distribution are considered, $\mathcal{N}(\mu_a = (0.2, 0.65), \Sigma_a = 0.01 I_2)$ for the class label a, $\mathcal{N}(\mu_b = (0.5, 0.9), \Sigma_b = 0.01 I_2)$ for the class label b and $\mathcal{N}(\mu_c = (0.8, 0.6), \Sigma_c = 0.01 I_2)$ for the class label c. In addition, a testing dataset of 50 samples for each label are generated using the same bivariate Gaussian distributions with a Gaussian noise $\mathcal{N}(\mu = (0,0), \Sigma = 0.001 I_2)$. In the end, we have a dataset of 1500 training data and 150 test data. First, nine point prediction classifiers are trained and tested on these data. The point prediction classifiers considered are the naive Bayes (*nbc*), the k-Nearest Neighbour (*knn*), the evidential k-Nearest Neighbour (*eknn*), the decision tree (*cart*), the random forest (*rfc*), linear discriminant analysis (*lda*), support vector machine (*svm*) and artificial neural networks (*ann*), the logistic classifier (*logistic*). The obtained accuracies are: logistic, ann: 94.67; svm, eknn: 95.33; and knn, nbc, rfc, lda, cart: 96. These classifiers are introduced here to detect the samples that are considered as “strange samples” in this paper, i.e., most point prediction classifiers fail to predict the true class of those samples. In the opposite case, the samples are considered as “usual samples”. This term will also be used, for a given classifier, to distinguish the samples for which the predicted class obtains a large probability, i.e., usual, from the others, i.e., strange.

4.1.1. The case of ndc classifier

The idea here for choosing the *ndc* hyper-parameter β is to avoid misclassification when the samples are strange and then predict a subset of classes for those samples. For the samples that are “usual”, the posterior probability of one of the classes is close to 1, thus the later class obtain the maximum utility regardless the value given to β (see Section 3.1). Consequently, it is more interesting to fix the value of β regarding the strange samples in the validation step. Indeed, the training data of 1500 samples is divided to 1200 samples for training and 300 (20%) samples for validation. The proposition of this paper is to consider a fictive probability distribution p^f where the first component is the mean of the maximal probabilities p^1 obtained for strange samples of the validation data set and the second component is the mean of the second maximal probabilities p^2, and so on. Thus, $p^f = (p^1, p^2, \ldots)$. To determine the strange samples a probability threshold is considered and it is fixed at 0.99 in this illustration. The value of β is considered as the threshold behind which if the samples are considered strange, we should predict the subset of the two first classes with maximal probabilities. In Proposition 3.1 this theoretical value corresponds to:

$$\beta_{ndc} = \sqrt{\frac{p^1 - p^2}{p^2}}. \tag{17}$$

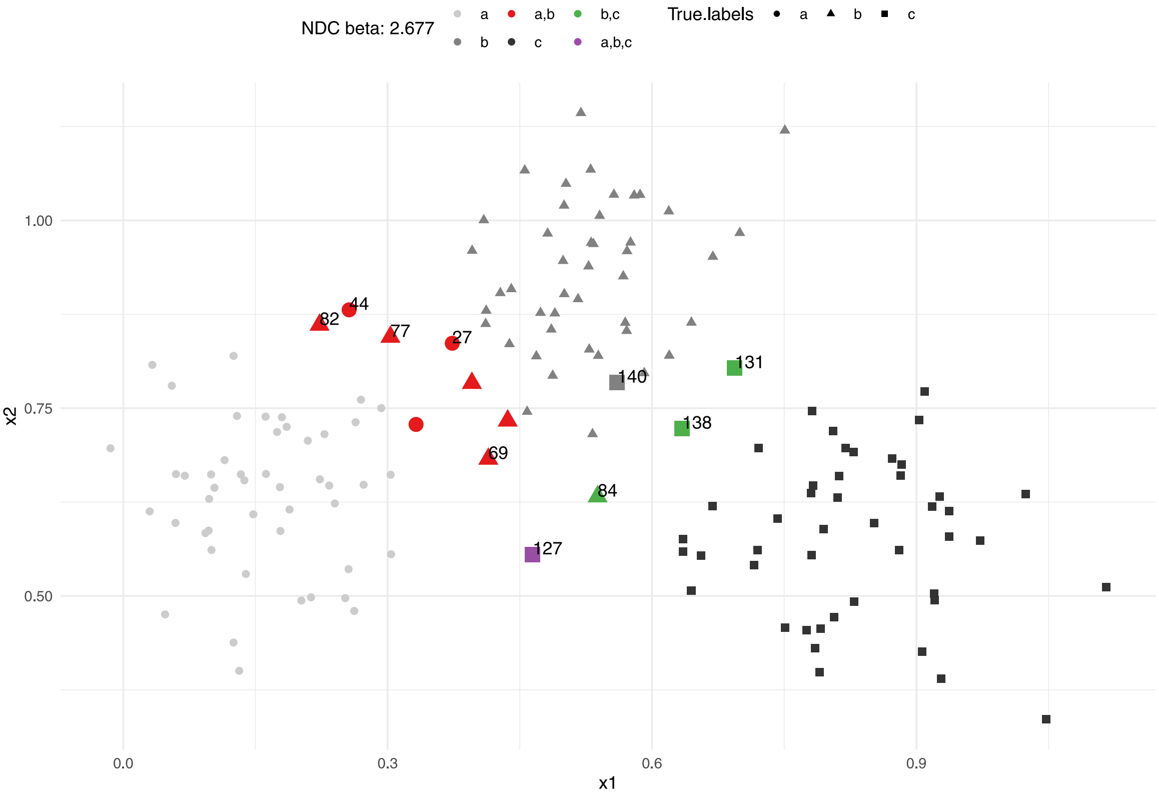


Fig. 3. The predictions obtained with *ndc*: a large size is given to the point symbols representing predictions that are errors or imprecise.

In Fig. 3, we present the predictions on the data set when $\beta_{ndc} = 2.677$ is determined as in Eq. (17) on the validation data. The samples that are considered strange for the point prediction classifiers are labelled by their number in the test dataset. Only the sample number 140 is misclassified in the predictions of ndc and only three “usual” samples have imprecise predictions. In general, among the ten strange samples, eight samples are predicted as subsets of two classes containing the true class and one as the whole set.

4.1.2. The case of eclair classifier

For the case of *eclair* classifier, we consider binary classifications “a against b”, “a against c” and “b against c”. We apply the same reasoning as in Section 4.1.1, the training data of is divided to 80% samples for training and 20% samples for validation. Here, also we consider only strange samples with the same mass threshold, i.e., 0.99. From Section 3.2, to avoid misclassification for strange samples β should be heigh enough to predict θ_{12} when ignorance is heigh. Let us denote m_{12} the average of $m(\theta_{12}|x)$ obtained for each strange sample in the validation data set. The proposed value of β is β_{eclair} that is the solution of the quadratic Eq. (16) with $m(\theta_{12}|x) = m_{12}$ and $m(\theta_1|x) = 2 \cdot (1 - m_{12})/3$. In Fig. 4, we can see that, for the case “a against b” ($\beta_{eclair} = 0.519$), we have one misclassification and four set-valued classification for the strange samples. For the case of “a against c” ($\beta_{eclair} = 0$), we have one misclassification and all the other strange samples are good predictions. While for the case “b against c”($\beta_{eclair} = 0.581$), we have one misclassification and one set-valued classification for the strange samples.

4.2. Illustration using fashion mnist data

In this section we propose to select the hyper-parameter β for the two classifiers *ndc* and *eclair* based on Eq. (10) for the modern version of mnist dataset [20], i.e. fashion mnist dataset [21]. These data are more difficult to handle compared to the original ones. Fashion-MNIST, a direct drop-in replacement for the original Y. Lecun’ MNIST dataset for benchmarking machine learning algorithms [20], is a dataset of Zalando’s article images [21] consisting of a training set of 60,000 examples and a test set of 10,000 examples. Each example is a 28 × 28 gray-scale image, associated with a label from 10 classes. Fig. 5 shows several images from this dataset where each class takes three-rows. As one can see in Figs. 6 and 7 (taken from [22]) from the visualization of datasets that makes comparison between mnist and fashion mnist [21, 22], fashion mnist dataset seems to be more challenging while for mnist dataset, classes are clearly separated.

Note that this illustration is presented only for the ndc classifier. First, we apply a grid search optimization to select the hyper-parameter β for ndc classifier. The grid search method is based on the objective function u_{65} as it is better suited to quantify the compromise between precision and cautiousness. Fig. 8 shows the performances obtained for the selected grid within the interval $[0, 3]$ on the validation data sets, i.e., 20% of training data, $\beta = 0.857$ gives the optimal mean performance at 0.954.

The fashion mnist data set has 10 classes, so it is more difficult to fix the probability threshold for strange samples compared to the simulated dated in Section 4.1. Thus, to calculate the different values of β in the same way as in Section 4.1.1, we considered several threshold of probabilities and represent the number of predicted subsets for each value β_1 calculated as in Eq. (11) on the validation data. Let recall that the number of data test is 10000. Moreover, the point prediction classifier used regarding the data nature to learn the posterior probabilities is the artificial neural networks (sequential model). Fig. 9 gives the parameter details of the employed architecture. Figures from 10 to 17 give the number of times a subset of 2, 3, 4, 5, 6 and 10 are predicted related of the values of β given in the legend. Obviously when changing the threshold of probabilities for strange examples, the values of β_1 change. In Fig. 10, the value $\beta_1 = 0.511$ corresponds to the β beyond which the utility of predicting a subset of size 2 is higher than the utility of prediction a single class; the value $\beta_1 = 1.004$ corresponds

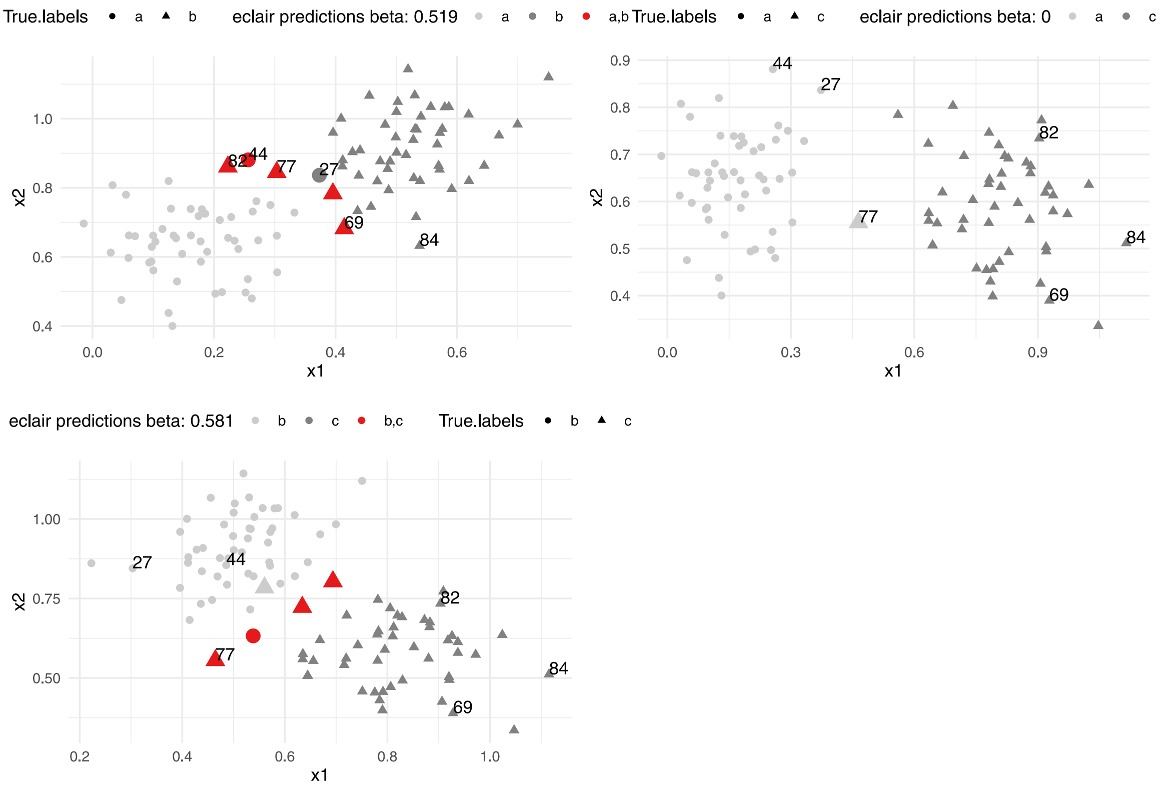


Fig. 4. The predictions obtained with *eclair*: a large size is given to the point symbols representing predictions that are errors or imprecise.

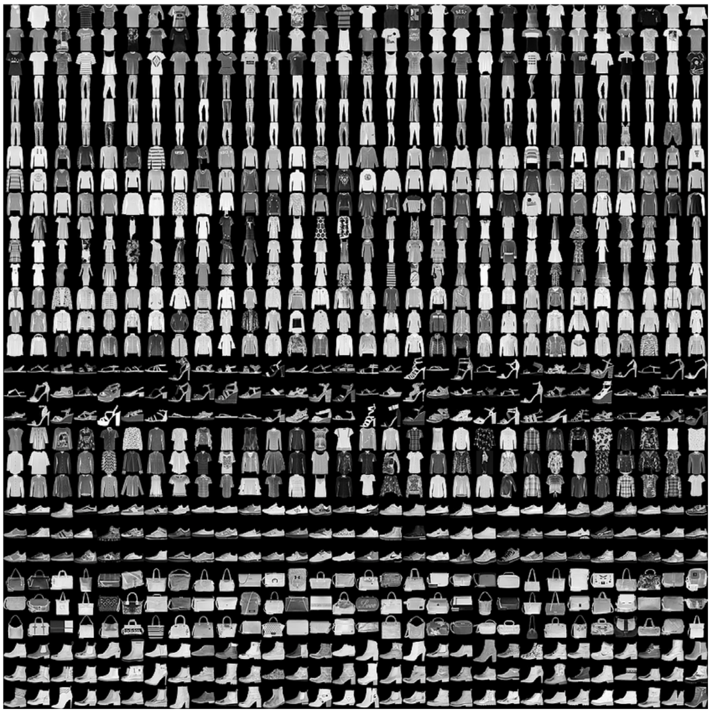


Fig. 5. Fashion-MNIST samples (by Zalando, MIT License).
The classes are: 1=‘T-shirt/top’, 2=‘Trouser’, 3=‘Pullover’, 4=‘Dress’, 5=‘Coat’, 6=‘Sandal’, 7=‘Shirt’, 8=‘Sneaker’, 9=‘Bag’, 10=‘Ankle boot’.

to the β beyond which the utility of predicting a subset of size 3 is higher than the utility of prediction a subset of size 2; and so on. One can see that only subsets of size 2, 3, 4 and 5 are predicted, the rest of the prediction are single classes. Furthermore, we can observe a slight increase of the predictions of 2, 3, 4 and 5 classes when the probability threshold increase. This is due to the mean of probabilities

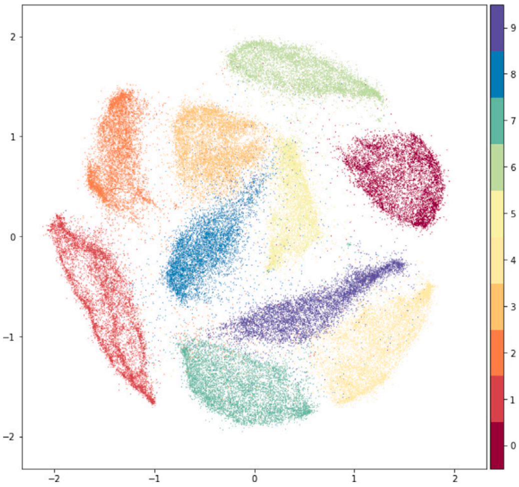


Fig. 6. mnist data clustering.

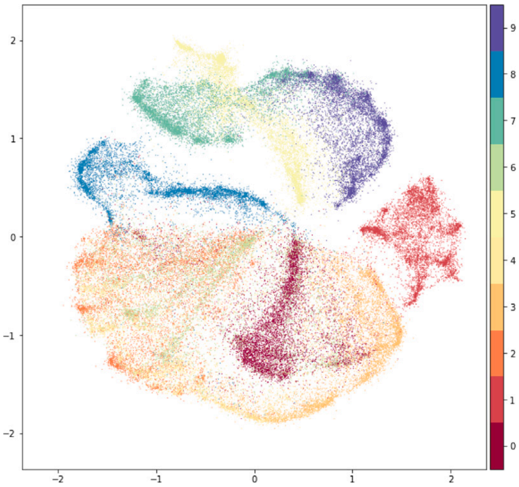


Fig. 7. Fashion mnist data clustering.

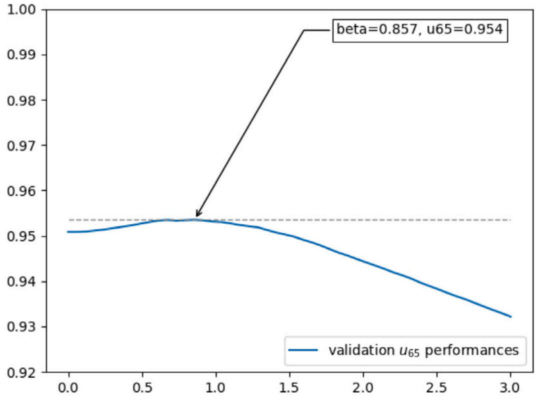
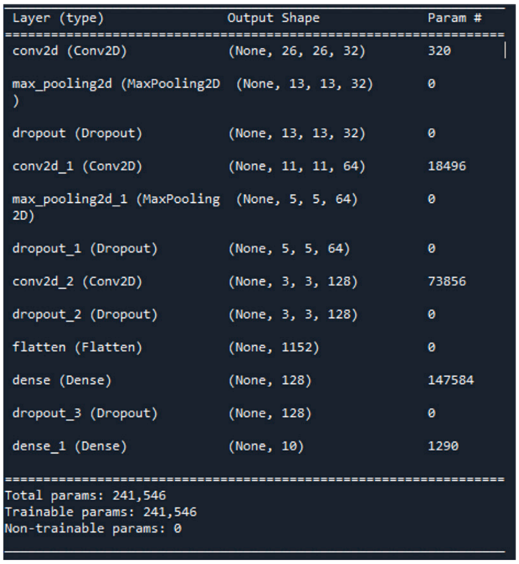


Fig. 8. β hyper-parameter grid search optimization for *ndc*.



```
Layer (type)                Output Shape              Param #
=================================================================
conv2d (Conv2D)             (None, 26, 26, 32)        320

max_pooling2d (MaxPooling2D  (None, 13, 13, 32)       0
)

dropout (Dropout)           (None, 13, 13, 32)        0

conv2d_1 (Conv2D)           (None, 11, 11, 64)        18496

max_pooling2d_1 (MaxPooling  (None, 5, 5, 64)         0
2D)

dropout_1 (Dropout)         (None, 5, 5, 64)          0

conv2d_2 (Conv2D)           (None, 3, 3, 128)         73856

dropout_2 (Dropout)         (None, 3, 3, 128)         0

flatten (Flatten)           (None, 1152)              0

dense (Dense)               (None, 128)               147584

dropout_3 (Dropout)         (None, 128)               0

dense_1 (Dense)             (None, 10)                1290

=================================================================
Total params: 241,546
Trainable params: 241,546
Non-trainable params: 0
_________________________________________________________________
```

Fig. 9. The neural networks architecture.

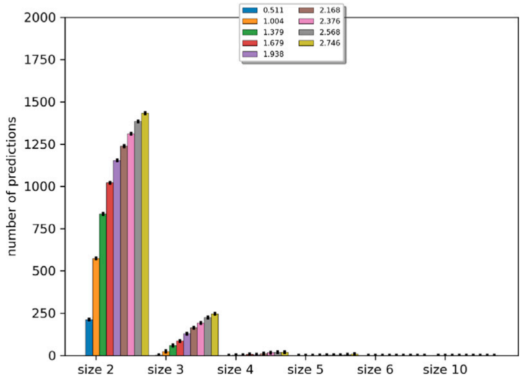


Fig. 10. Threshold fixed at $p = 0.55$.

Table 3
u_{65} performances for difference values of β and p on the test data of fashion mnist.

	$p = 0.55$	$p = 0.6$	$p = 0.65$	$p = 0.7$	$p = 0.75$	Grid search
u_{65}	0.9241	0.9245	0.9249	0.9245	0.9244	0.9244
β_1	1.379	1.118	1.224	1.339	0.803	0.857

that are influenced by the values of large probabilities of "usual" samples.

Table 3 shows the best u_{65} performance measure for each probability threshold. One can see also, in the last column, the u_{65} performance measure obtained using the optimal hyper-parameters tuned with grid search method. As we can except for $p = 0.55$, the results obtained with our proposition are better than grid search one.

5. Related works

The closest work, in principle, to that of the proposal of this article, is the one of conformal prediction. Conformal prediction is designed to perform label predictions successively, each one begin revealed before the next is predicted [11–13], but it is also adapted to classical prediction task as for regression [14]. The general principal is the following:

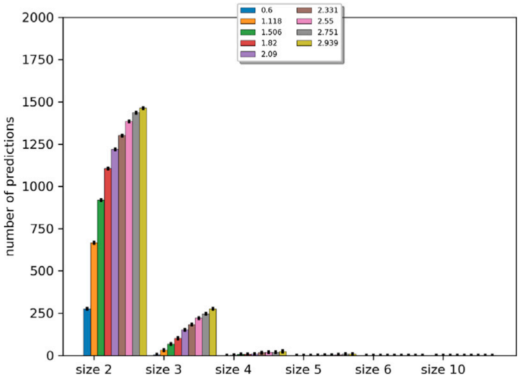


Fig. 11. Threshold fixed at $p = 0.6$.

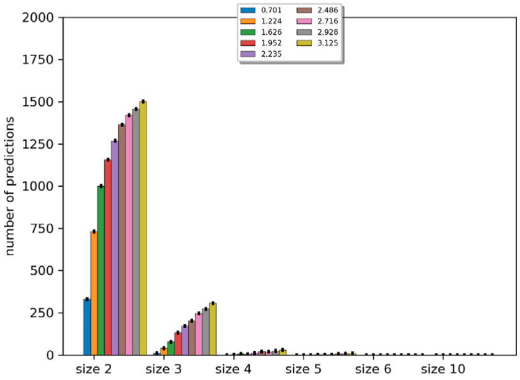


Fig. 12. Threshold fixed at $p = 0.65$.

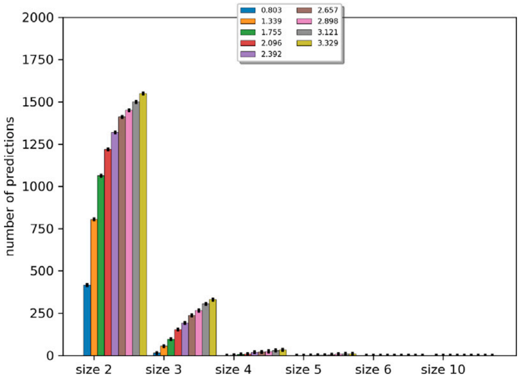


Fig. 13. Threshold fixed at $p = 0.7$.

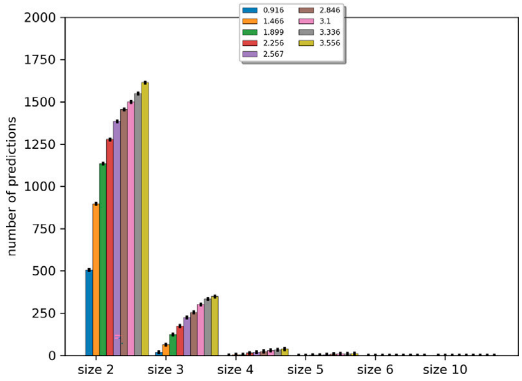


Fig. 14. Threshold fixed at $p = 0.75$.

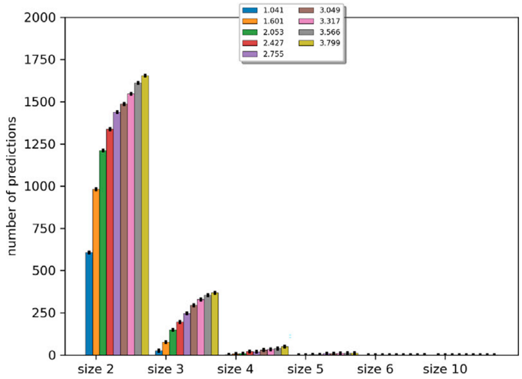


Fig. 15. Threshold fixed at $p = 0.8$.

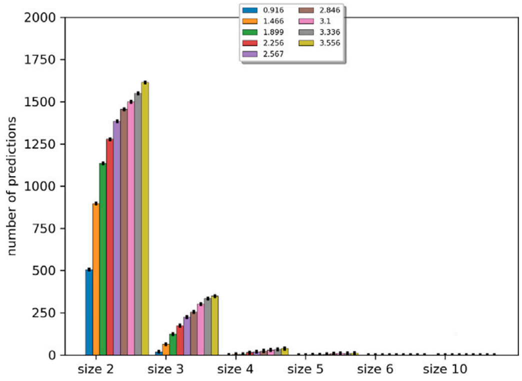


Fig. 16. Threshold fixed at $p = 0.85$.

(1) training data are divided to training part and calibration/validation part; (2) a classifier/regressor δ is learnt using the training part; (3) non-conformity (strangeness) score is computed on the calibration part by comparing the predictions of δ to the true labels; (4) in the same way as in 3-, each potential class y of a sample x in the test data is associated a non-conformity score α_y related the prediction of δ; (5) then the p-value is defined for the potential class y of x as the portion of the calibration data and x that have a non-conformity scores greater than α_y; (6) for a given small positive value ϵ (1% or 5%), the predictive region output (or the set-valued prediction) is the subset: $\{y : p(y) > \epsilon\}$. The common point between the two classifiers is the fact of building the predictions for the new samples on the basis of the non-conformity or the strangeness of some (or all) the samples of the calibration data compared to what was learned by the classifiers. While the difference lies in how the non-conformity scores are calculated on the one hand, and how the predictions are performed on the other hand. Indeed, in our proposal, non-conformity score is calculated with respect to the certainty of the prediction, i.e., how closely the sample resembles other training samples regardless the true class of the calibration samples. Concerning the set-valued predictions conformal prediction is based on the concept of confidence level which well established in statistics, while in our proposal predictions are based on a compromise through a subjective utility which compares the subsets of classes. Moreover, Table 4 shows the results obtained by the conformal predictions for two different confidence levels (0.95% and 0.99%) for the fashion mnist data. As one can see in the comparison with the result shown in Table 3, our proposition obtained better u_{65} scores. But with a high confidence level, conformal prediction shows a very high score for "impr. acc" at

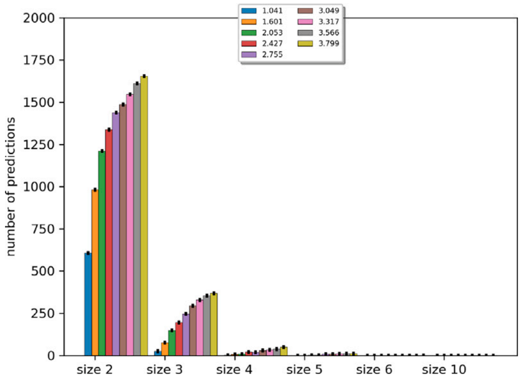


Fig. 17. Threshold fixed at $p = 0.9$.

Table 4
Conformal predictions for the test data of fashion mnist.

	acc	u_{50}	u_{65}	u_{80}	Impr. acc
$\epsilon = 0.05$	0.6785	0.792	0.83	0.868	0.947
$\epsilon = 0.01$	0.4817	0.6642	0.729	0.7945	0.9888

the expense of a very low accuracy score. The decision rule strategies have different complexity and it is challenging as the set of alternatives size is 2^n, $n = |\Theta|$. The *ndc* classifier has the lowest optimized computational complexity, i.e., at worst $O(n)$ [5]. However, the computational complexity is challenging for conformal prediction (see [14] for more details) and for *eclair* classifiers (see [10] for more details). Indeed, for conformal predictions, the non-conformity scores are calculated using the nearest neighbours of the calibration or test samples in the training data [11]. Regarding *eclair* classifier, as for any approach representing imprecision in the data, the computational complexity can be very high. Indeed, the computational complexity of the reasoning step of the *eclair* classifiers becomes very high, i.e., $O(2^{2n})$, at worst. One can find in [10] some optimizations to overcame the problem. For example, by selecting the relevant candidate subsets for prediction to a subset of 2^{Θ}.

6. Conclusion

In this paper we are interested in the set-valued classification. Especially, we focus on the study of the parameter β involved in the utility function used in the decision step of two set-valued classifiers. More precisely, we studied the predicted subsets depending on this hyper-parameter. In addition to theoretical propositions, we give practical method to control the size of the predicted subset in machine learning applications. While trying to remain very efficient on point prediction task, set-valued classifiers have the challenge of making machine learning methods more trustworthy, especially in the presence of imperfect data. The decision-maker who knows well his data could better control, using the proposal of this article, the size of the predictions by fixing the suited value for β. As a perspective, we intend in our next work to provide a demonstration that relies on theoretical foundations and statistical hypotheses to support the choices of the different thresholds that are experimentally set in our illustrations. We will also try to handle the merging part of the “one against all” solution suggested in the *eclair* part. Indeed, we obtain different values of β_3 for each pair comparisons which make the merging part of the pair comparisons difficult to handle.

CRediT authorship contribution statement

Abdelhak Imoussaten: Study conception and design, Data collection, Analysis and interpretation of results, Writing – original draft.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The information about the data are given in the paper.

References

[1] Jacquin L, Imoussaten A, Trousset F, Perrin D, Montmain J. Control of waste fragment sorting process based on MIR imaging coupled with cautious classification. Resour Conserv Recy 2021;168:105258.
[2] Zaffalon M. A credal approach to naive classification. In: ISIPTA, vol. 99. 1999, p. 405–14.
[3] Zaffalon M. Statistical inference of the naive credal classifier. In: ISIPTA, vol. 1. 2001, p. 384–93.
[4] Troffaes MC. Decision making under uncertainty using imprecise probabilities. Int J Approx Reason 2007;45(1):17–29.
[5] Coz JJd, Díez J, Bahamonde A. Learning nondeterministic classifiers. J Mach Learn Res 2009;10(Oct):2273–93.
[6] Abellán J, Moral S. Building classification trees using the total uncertainty criterion. Int J Intell Syst 2003;18(12):1215–25.
[7] Abellan J, Masegosa AR. Imprecise classification with credal decision trees. Int J Uncertain Fuzziness Knowl-Based Syst 2012;20(05):763–87.
[8] Ma L, Denoeux T. Partial classification in the belief function framework. Knowl-Based Syst 2021;106742.
[9] Jacquin L, Imoussaten A, Trousset F, Montmain J, Perrin D. Evidential classification of incomplete data via imprecise relabelling: Application to plastic sorting. In: Ben Amor N, Quost B, Theobald M, editors. Scalable uncertainty management. Cham: Springer International Publishing; 2019, p. 122–35.
[10] Imoussaten A, Jacquin L. Cautious classification based on belief functions theory and imprecise relabelling. Internat J Approx Reason 2022;142:130–46.
[11] Shafer G, Vovk V. A tutorial on conformal prediction. J Mach Learn Res 2008;9(Mar):371–421.
[12] Vovk V, Gammerman A, Shafer G. Conformal prediction. Algorithmic Learn Random World 2005;17–51.
[13] Gammerman A, Vovk V, Vapnik V. Learning by transduction. 2013, arXiv preprint arXiv:1301.7375.
[14] Papadopoulos H, Proedrou K, Vovk V, Gammerman A. Inductive confidence machines for regression. In: Machine learning: ECML 2002: 13th European conference on machine learning Helsinki, Finland, August 19–23, 2002 proceedings 13. Springer; 2002, p. 345–56.
[15] Couso I, Sánchez L. Machine learning models, epistemic set-valued data and generalized loss functions: an encompassing approach. Inform Sci 2016;358:129–50.
[16] Sanchez L, Couso I. A framework for learning fuzzy rule-based models with epistemic set-valued data and generalized loss functions. Internat J Approx Reason 2018;92:321–39.
[17] Yang G, Destercke S, Masson M-H. The costs of indeterminacy: How to determine them? IEEE Trans Cybern 2016;47(12):4316–27.
[18] Tsoumakas G, Vlahavas I. Random k-labelsets: An ensemble method for multilabel classification. In: European conference on machine learning. Springer; 2007, p. 406–17.
[19] Zaffalon M, Corani G, Mauá D. Evaluating credal classifiers by utility-discounted predictive accuracy. Internat J Approx Reason 2012;53(8):1282–301.
[20] LeCun Y. The MNIST database of handwritten digits. 1998, http://yann.lecun.com/exdb/mnist/.
[21] Xiao H, Rasul K, Vollgraf R. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. 2017, arXiv preprint arXiv:1708.07747.
[22] Agrawal A, Ali A, Boyd S. Minimum-distortion embedding. 2021, arXiv.