Choosing the Decision Hyper-parameter for Some Cautious Classifiers

Abdelhak Imoussaten

Choosing the decision hyper-parameter for some cautious classifiers

Abdelhak Imoussaten[0000−0002−1292−2681]

EuroMov Digital Health in Motion, Univ Montpellier, IMT Mines Ales, Ales, France
abdelhak.imoussaten@mines-ales.fr

Abstract. In some sensitive domains where data imperfections are present, standard classification techniques reach their limits. To avoid misclassification that has serious consequences, recent works propose cautious classification algorithms to handle the problem. Despite of the presence of uncertainty, a point prediction classifier is forced to decide which single class to associate to a sample. In such a case, 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 override relevance 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 of the classifier algorithm 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 eclair (evidential classifier based on imprecise relabelling) proposed within the framework of belief functions. The theoretical aim of the mentioned parameter is to control the size of predictions for both classifiers. This paper proposes to study this parameter in order to select the "best" value in a classification task. First the gain for each prediction candidate is studied related to the values of the hyper-parameter. In the illustration section, we propose a method to choose this hyper-parameter base on the training data and we show the classification results on randomly generated data and we present some comparisons with two other imprecise classifiers on 11 UCI datasets based on five measures of imprecise classification performances used in the state of the art.

Keywords: Cautious classification · Imprecise classification · Belief functions · Supervised machine learning.

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 [6], e.g., in medical diagnosis applications when a classifier is involved to detect early-stage cancer. In such applications cautiousness is necessary when imperfect data are present. This leads some recent works to focus on cautious classification. Among the existing cautious classifiers, we focus, in

this paper, on those providing a subset of candidate class labels to a new sample to classify and we called them *imprecise classifiers*. Some of them, as the non-deterministic classifier (*ndc*) [3], use the posterior probability when it is known and provide the subset of classes, that minimize/maximize a risk/utility function, as prediction (see Subsection 2.2 for more details). Other approaches, as the *Naive Credal Classifier* (*ncc*) [12] [13] proposed in the framework of imprecise probability, are based on a dominance relation defined on the set of classes using the *credal* set representing the imprecision and uncertainty about the true class label of a sample. Then the subset of the non-dominated classes is considered as the prediction for the sample. The imprecise classifiers proposed within the framework of belief functions utilise the mass function when it is known and a decision procedure. In [7], it is proposed to generalize the utility matrix to the subsets of classes by aggregating the single utilities that are considered as known. The approach in [8] uses the interval dominance approach where the intervals are represented by the values of belief and plausibility functions obtained of each class. In [5] [4], the evidential classifier based on imprecise relabelling (*eclair*) uses a generalisation of the gain function proposed in [3] to the case of belief functions framework. An imprecise classifier proposes the appropriate subset of candidate classes that can be assigned to the sample in the presence of imperfect information. But cautiousness should not override relevance 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 not relevant. On the other hand, a classifier that predicts always a single class for difficult samples is relevant when the prediction is good but it is not cautious. Most of imprecise classifiers cannot control this trade-off except *ndc* and *eclair*. Indeed, the gain function implemented in the decision step of both classifiers *ndc* and *eclair* has an hyper parameter β that is used to control the trade-off between relevance and cautiousness. This hyper parameter is considered as a user-modifiable parameter for the use of these two 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 in 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 a suggestion 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 parameter value on the prediction of the two classifiers when faced to difficult samples, i.e., to which the standard classifiers failed to predict the true class labels. While in the second experiment part, we present some comparison of the *ndc* classifier tuned using our proposition with other imprecise classifiers of the state of the art conduct on 11 UCI data and based on five measures from the state of the art that are usually used to compare imprecise classification performances. The paper is organised as follows. In the second section, the reminders about the decision step in the classifiers *eclair* and *ndc* and the measures of imprecise classification performances are given. The third section presents a study of the expected gain

function introduced in the decision step of the two classifiers. Finally, the fourth section presents the experiment results.

2 Reminders and notations

The imprecise 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 to classify. We focus in this paper on the decision step of those two classifiers that involves these two functions and a gain 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. Finally, five measures from the state of the art used to evaluate the imprecise classification performances are presented. To simplify notations, we adopt the following notations for the subsets in the rest of the paper: $\theta_i := \{\theta_i\}$, $\theta_{ij} := \{\theta_i, \theta_j\}$.

2.1 F_β measure

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 introduced in information retrieval and classification to imprecise classification. In the context of binary point prediction for classification, the F_β score is defined as:

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

where $\text{precision} = \frac{true\ positive}{true\ positive + false\ positive}$ and $\text{recall} = \frac{true\ positive}{true\ positive + false\ negative}$ are two known performance measures in information retrieval and machine learning.

2.2 The decision step in *ndc*

The principle of *ndc* is very simple, a posterior probability is determined using a classification method for point prediction and then a decision rule is applied to determine the imprecise prediction. This subsection presents the decision rule applied in the decision step. The decision step with *ndc* consists in providing for a sample $\mathbf{x}$ a subset of classes as prediction, i.e., precise predictions are given as singletons, by considering as input the posterior probability $p(.|\mathbf{x})$. The predicted subset of classes is the one maximizing the expected gain where the gain associated to each subset of classes is defined using the F_β measure. More precisely, let us consider a set of n class labels $\Theta = \{\theta_1, \ldots, \theta_n\}$. Each subset of candidate classes $A \subseteq \Theta$ is evaluated as the good prediction for $\mathbf{x}$ using the F_β measure as follows:

$$F_\beta(A, \mathbf{x}, \theta) = \frac{(1+\beta^2) \cdot \mathbb{1}_A(\theta)}{\beta^2 + |A|}. \tag{2}$$

The quantity $F_\beta(A, \mathbf{x}, \theta)$ is interpreted as the gain obtained when predicting the subset of class labels A for the sample $\mathbf{x}$ when its true class label is θ. The Formula in (2) is analogue to the one in (1) where the quantities *precision* and *recall* are redefined as $\text{precision}(A) = \frac{\mathbb{1}_A(\theta)}{nb\,of\,classes\,in A}$ and $\text{recall}(A) = \mathbb{1}_A(\theta)$ but do not have the same meaning. Indeed, in (1) the case of binary classification, the two measures are quantified related to a data test set while in the case of imprecise classification the two measures are quantified related to a subset of classes that is a potential prediction. We can note that when the values of β are close to 0, $F_\beta(A, \mathbf{x}, \theta)$ becomes close to $\text{precision}(A)$ thus the size of A is disadvantageous. On the other hand, when β is high, $F_\beta(A, \mathbf{x}, \theta)$ becomes close to $\text{recall}(A)$ and in this case the size of A is an advantage. Let us suppose that a posterior probability distribution $p(.|\mathbf{x})$ is known for the sample $\mathbf{x}$, the non-deterministic classifier *ndc* predicts for x the subset of candidate classes that maximize the expected gain function $u_\beta(., p(.|\mathbf{x}))$ defined as:

$$u_\beta(A, p(.|\mathbf{x})) = \sum_{i=1}^{n} F_\beta(A, \mathbf{x}, \theta_i) \cdot p(\theta_i|\mathbf{x}). \tag{3}$$

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

$$\delta_{ndc}(\mathbf{x}) = arg \max_{A \subseteq \Theta} u_\beta(A, p(.|\mathbf{x})). \tag{4}$$

2.3 The decision step in *eclair*

The decision step with *eclair* consists in providing for a sample $\mathbf{x}$ a subset of classes as prediction, by considering as input the posterior mass function $m(.|\mathbf{x})$. The predicted subset of classes is the one maximizing the expected gain where the gain associated to each subset of classes is defined using a generalisation of the formula (2) [5] [4]. The main change is to consider the general case where the available information about the true class of a sample can be partial in the form of a subset $B \subseteq \Theta$. It is the case, for example, when data are coarse [2] [9]. This leads to the new gain function defined as follows:

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

The quantity $F_\beta(A, \mathbf{x}, B)$ is interpreted as the gain obtained when predicting the subset of class labels A for the sample $\mathbf{x}$ when its true class label is partially known and represented by a subset of classes B. In this case, the precision and recall analogue quantities of ones presented in (1) become: $\text{precision}(A) = \frac{|A \cap B|}{nb\,of\,classes\,in A}$ and $\text{recall}(A) = \frac{|A \cap B|}{nb\,of\,classes\,in B}$.

Let us suppose that a posterior mass function $m(.|\mathbf{x})$ is known for the sample $\mathbf{x}$, the *eclair* classifier predicts for $\mathbf{x}$ the subset of candidate classes that maximize the expected gain function $u_\beta(., m(.|\mathbf{x}))$ defined as:

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

Finally, the predicted subset $\delta_{eclair}(\mathbf{x})$ for $\mathbf{x}$ using the classifier *eclair* is given as:

$$\delta_{eclair}(\mathbf{x}) = arg \max_{A \subseteq \Theta} u_\beta(A, m(.|\mathbf{x})). \tag{7}$$

2.4 Evaluation measures for the imprecise classifiers

When evaluating an imprecise classifier one ensures that the predicted subset of classes 1) include the "true" class and 2) they are 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) [12],[1],[11]. 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* [10] 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 [14] for more details).

Let us consider a dataset of test samples $dst = (\boldsymbol{x}^l, \theta^l)_{1 \leq l \leq M}$ where $\boldsymbol{x}^l \in \mathcal{X}$ and $\theta^l \in \Theta$ and an imprecise classifier δ_{ic}. The five following measures are proposed to evaluate the performance of imprecise classification and applied to the classifier δ_{ic} and the test data dst:

- the *classical accuracy*:

$$accuracy(\delta_{ic}, dst) = \frac{1}{M} \sum_{l=1}^{M} \mathbb{1}_{\{\theta^l\}}(\delta_{ic}(\boldsymbol{x}^l)).$$

- the *imprecise accuracy* (imprAcc):

$$imprAcc(\delta_{ic}, dst) = \frac{1}{M} \sum_{l=1}^{M} \mathbb{1}_{\delta_{ic}(\boldsymbol{x}^l)}(\theta^l).$$

- the *discounted accuracy* (discAcc) corresponds to the function $g(z) = z$ [10]:

$$discAcc(\delta_{ic}, dst) = \frac{1}{M} \sum_{l=1}^{M} \frac{\mathbb{1}_{\delta_{ic}(\boldsymbol{x}^l)}(\theta^l)}{|\delta_{ic}(\boldsymbol{x}^l)|},$$

  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 $g(z) = -0.6 \cdot z^2 + 1.6 \cdot z$ [14]:

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

– The u_{80} measure that corresponds to the function $g(z) = -1.2 \cdot z^2 \ + \ 2.2 \cdot z$ [14]:

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

3 The expected gains related to β

3.1 The case of ndc

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

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

In addition, let us consider the situation where the class θ_i is the most likely class of $\mathbf{x}$ and some times the class θ_i is confused with the class θ_j, $j \neq i$ due to data imperfection. The Propositions 1 and 2 give some results concerning the predicted subset of classes for $\mathbf{x}$ from the three options θ_i, θ_{ij} and Θ.

Proposition 1. *Let suppose that* $p(\theta_i|\boldsymbol{x}) > p(\theta|\boldsymbol{x})$, $\forall \theta \in \Theta \setminus \theta_i$.

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

$$\begin{cases} u(\beta, \theta_{ij}, p(.|\boldsymbol{x})) \leq u(\beta, \theta_i, p(.|\boldsymbol{x})) & \text{if } \beta \leq \beta_1 \\ u(\beta, \theta_{ij}, p(.|\boldsymbol{x})) > u(\beta, \theta_i, p(.|\boldsymbol{x})) & \text{if } \beta > \beta_1. \end{cases} \tag{9}$$

Elsewhere $u(\beta, \Theta, p(.|\boldsymbol{x})) < u(\beta, \theta_{ij}, p(.|\boldsymbol{x}))$, $\forall \beta \geq 0$.

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

$$u(\beta, \theta_i, p(.|\mathbf{x})) = p(\theta_i|\mathbf{x}).$$

and

$$u(\beta, \theta_{ij}, p(.|\mathbf{x})) = \frac{1 + \beta^2}{2 + \beta^2} \cdot [p(\theta_i|\mathbf{x}) + p(\theta_j|\mathbf{x})].$$

On the one hand, the function $u(., \theta_{ij}, p(.|\mathbf{x}))$ increases related to β. Thus $u(\beta, \theta_{ij}, p(.|\mathbf{x})) \geq \frac{1}{2}(p(\theta_i|\mathbf{x}) + p(\theta_j|\mathbf{x}))$, for all $\beta \geq 0$. On the other hand, $p(\theta_i|\mathbf{x}) > p(\theta_j|\mathbf{x})$ then $p(\theta_i|\mathbf{x}) > \frac{1}{2}(p(\theta_i|\mathbf{x}) + p(\theta_j|\mathbf{x}))$. So, $u(., \theta_{ij}, p(.|\mathbf{x}))$ intersects $u(., \theta_i, p(.|\mathbf{x}))$ at $\beta_1 \geq 0$ such that:

$$\frac{1 + \beta_1^2}{2 + \beta_1^2} \cdot [p(\theta_i|\mathbf{x}) + p(\theta_j|\mathbf{x})] = p(\theta_i|\mathbf{x}).$$

It comes:

$$\beta_1 = \sqrt{\frac{p(\theta_i|\mathbf{x}) - p(\theta_j|\mathbf{x})}{p(\theta_j|\mathbf{x})}}.$$

■

Proposition 2. *Let suppose that* $p(\theta_i|\boldsymbol{x}) > p(\theta|\boldsymbol{x})$, $\forall \theta \in \Theta \setminus \theta_i$. *If* $\mathbb{P}(\theta_{i,j}|\boldsymbol{x}) \in [\frac{2}{3}, 1[$ *then it exists* $\beta_2 > 0$ *such that:*

$$\begin{cases} u(\beta, \Theta, p(.|\boldsymbol{x})) \leq u(\beta, \theta_{ij}, p(.|\boldsymbol{x})) & \text{if } \beta \leq \beta_2 \\ u(\beta, \Theta, p(.|\boldsymbol{x})) > u(\beta, \theta_{ij}, p(.|\boldsymbol{x})) & \text{if } \beta > \beta_2. \end{cases} \tag{10}$$

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

$$u(\beta, \Theta, p(.|\mathbf{x})) = \frac{1+\beta^2}{3+\beta^2},$$

and

$$u(\beta, \Theta, p(.|\mathbf{x})) - u(\beta, \theta_{ij}, p(.|\mathbf{x})) = \frac{(1+\beta^2)\cdot(2-3\cdot\mathbb{P}(\theta_{i,j})+(1-\mathbb{P}(\theta_{i,j})\cdot\beta^2))}{(3+\beta^2)\cdot(2+\beta^2)}$$

where $\mathbb{P}(\theta_{i,j}|\mathbf{x}) = p(\theta_i|\mathbf{x}) + p(\theta_j|\mathbf{x})]$. If $\mathbb{P}(\theta_{i,j}|\mathbf{x}) < \frac{2}{3}$, then $u(\beta, \Theta, p(.|\mathbf{x})) > u(\beta, \theta_{ij}, p(.|\mathbf{x}))$, $\forall \beta \geq 0$. Else, if $\mathbb{P}(\theta_{i,j}|\mathbf{x}) = 1$, then $u(\beta, \Theta, p(.|\mathbf{x})) = \frac{1+\beta^2}{3+\beta^2} < \frac{1+\beta^2}{2+\beta^2} = u(\beta, \theta_{i,j}, p(.|\mathbf{x}))$, $\forall \beta \geq 0$. Otherwise, let us consider the following value $\beta^* \geq 0$ such that:

$$\beta^{*2} = \frac{3\ \mathbb{P}(\theta_{i,j}|\mathbf{x}) - 2}{1 - \mathbb{P}(\theta_{i,j}|\mathbf{x})}.$$

We can set

$$\beta_2 = \sqrt{\beta^{*2}}.$$

■

Example 1. Let us consider the following examples of four samples that obtain the posterior probabilities given in Figure 1. These distribution express several situation of sharing the masses between the three classes. For the first sample x_1 the mass is uniformly distributed on the classes; for x_2 the mass is totally given to the class θ_1; for x_3 the mass is uniformly distributed to θ_1 and θ_2; and for x_4 the mass distribution is as follows $p(\theta_3|\mathbf{x}_4) < p(\theta_1|\mathbf{x}_4) < p(\theta_2|\mathbf{x}_4)$. As one can see in figure 1, for the samples x_1, x_2 and x_3, Θ, θ_1, and $\theta_{1,2}$ are respectively the predictions as they maximize the expected gain regardless the value of β. In the case of x_4, the prediction depends on the value of the parameter β. Indeed, if $\beta < \beta_1 = \sqrt{\frac{p(\theta_1|\mathbf{x}_4)-p(\theta_2|\mathbf{x}_4)}{p(\theta_2|\mathbf{x}_4)}} = 0.5$, i.e., the value of β where the curves of $u(., \theta_1, \mathbf{x}_4)$ and $u(., \theta_{1,2}, \mathbf{x}_4)$ intersect, then θ_1 dominates all the other options. When $\beta_2 > \beta > \beta_1$ ($\beta_2 = \sqrt{\frac{3\ \mathbb{P}(\theta_{1,2}|\mathbf{x})-2}{1-\mathbb{P}(\theta_{1,2}|\mathbf{x})}} = 2.65$), 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

In this subsection, we consider that the posterior mass function of a sample $\mathbf{x}$ is known. We denote this mass function by $m(.|\mathbf{x}) : 2^\Theta \to [0, 1]$. In this case, the

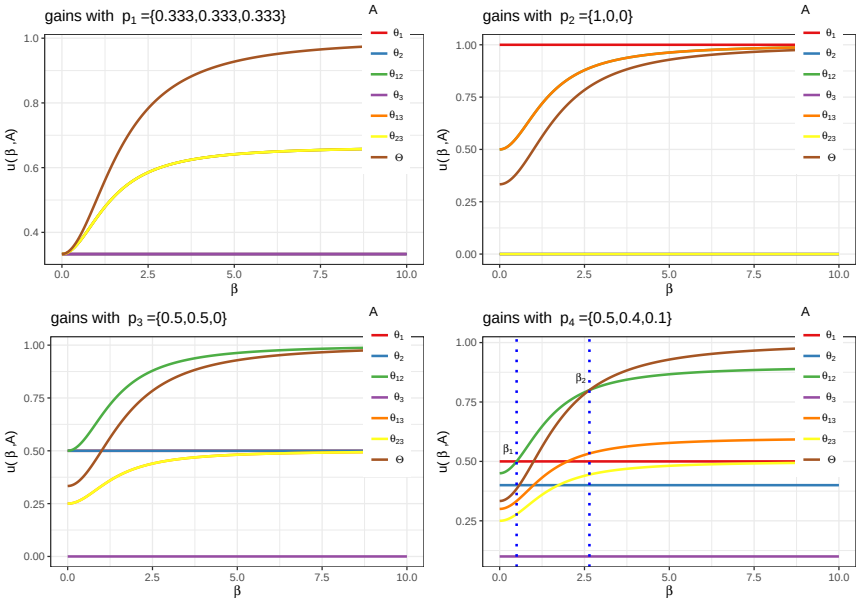

Fig. 1. the expected gain associated to the four posterior probabilities.

expected gain function used as the criterion to choose the subset of classes to associate to $\mathbf{x}$ is the following:

$$u(\beta, A, m(.|\mathbf{x})) = \sum_{B \subseteq \Theta} F_\beta(A, \mathbf{x}, B) \cdot m(B|\mathbf{x}) \tag{11}$$

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

Proposition 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:*

$$\begin{cases} u(\beta, \theta_{12}, m(.|\boldsymbol{x})) \leq u(\beta, \theta_1, m(.|\boldsymbol{x})) & \text{if } \beta \leq \beta_3 \\ u(\beta, \theta_{12}, m(.|\boldsymbol{x})) > u(\beta, \theta_1, m(.|\boldsymbol{x})) & \text{if } \beta > \beta_3 \end{cases} \tag{12}$$

Elsewhere, $u(\beta, \theta_{12}, m(.|\boldsymbol{x})) \geq u(\beta, \theta_1, m(.|\boldsymbol{x}))$, $\forall \beta \geq 0$.

Proof. In one hand, we have,

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

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

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

consequently $u(.,\theta_{12},m(.|\mathbf{x}))$ increases $\forall\beta\geq 0$ with $u(0,\theta_{12},m(.|\mathbf{x})))=\frac{1}{2}+\frac{m(\theta_{12}|\mathbf{x})}{2}$ and $\lim_{\beta\to+\infty}u(\beta,\theta_{12},m(.|\mathbf{x})))=1$. Obviously, if $u(0,\theta_1,m(.|\mathbf{x})))>u(0,\theta_{12},m(.|\mathbf{x})))$ then $u(.,\theta_1,m(.|\mathbf{x}))$ and $u(.,\theta_{12},m(.|\mathbf{x}))$ intersect, elsewhere $u(\beta,\theta_{12},m(.|\mathbf{x}))\geq u(\beta,\theta_1,m(.|\mathbf{x}))$, $\forall\beta\geq 0$. The inequality $u(0,\theta_1,m(.|\mathbf{x})))>u(0,\theta_{12},m(.|\mathbf{x})))$ corresponds to $m(\theta_1|\mathbf{x})+m(\theta_{12}|\mathbf{x})>\frac{1}{2}+\frac{m(\theta_{12}|\mathbf{x})}{2}$ which is verified when $m(\theta_1|\mathbf{x})>m(\theta_2|\mathbf{x})$. Finally, β_3 is the solution of $u(\beta,\theta_1,m(.|\mathbf{x}))=u(\beta,\theta_{12},m(.|\mathbf{x}))$ which corresponds to the solution of Equation (13):

$$m(\theta_1|\mathbf{x})+\frac{1+\beta^2}{1+2\,\beta^2}m(\theta_{12}|\mathbf{x})=\frac{1+\beta^2}{2+\beta^2}+\frac{1}{2+\beta^2}m(\theta_{12}|\mathbf{x}). \tag{13}$$

■

Remark 1. Note that when m is a Bayesian mass function, we have the Equation (13) giving β_3 that becomes: $m(\theta_1|\mathbf{x})=\frac{1+\beta_3^2}{2+\beta_3^2}$, which corresponds to

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

Example 2. To illustrate the different situations, we consider six mass functions (see Figure 2). Figure 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 higher the mass of ignorance is, the smaller β_3 becomes.

4 Illustration

In this section we present the illustration of the performances of the classifiers *ndc* and *eclair* using generated data and then we present the comparisons of the *ndc* classifier tuned using our proposition with other imprecise classifiers on the UCI data based on the five measures presented in Subsection 2.4.

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, $\mathcal{N}(\mu_a=(0.2,0.65),\Sigma_a=0.01I_2)$ for the class label a, $\mathcal{N}(\mu_b=(0.5,0.9),\Sigma_b=0.01I_2)$ for the class label b and $\mathcal{N}(\mu_c=(0.8,0.6),\Sigma_c=0.01I_2)$ for the class label c. In addition, a testing dataset of 50 samples for

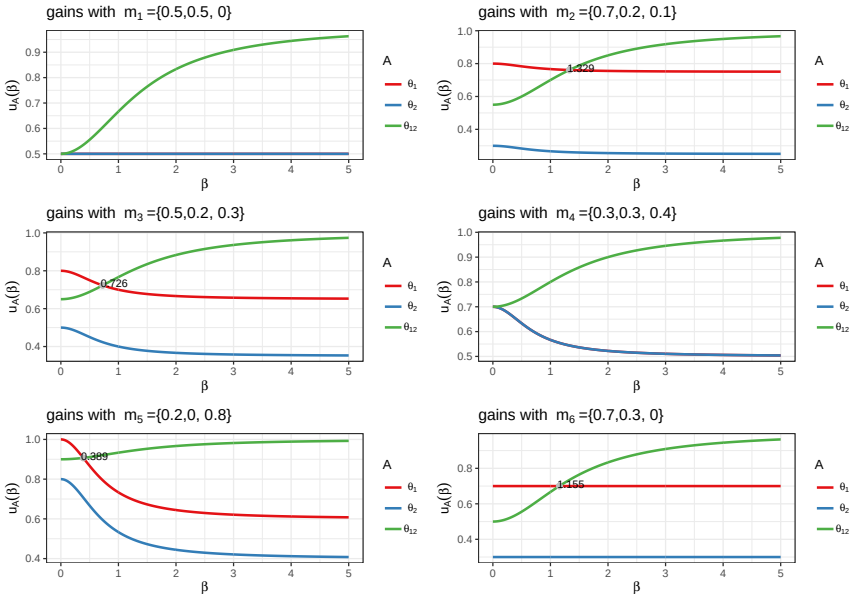

Fig. 2. the gain function for some examples of masses

each label are generated using the same bivariate Gaussian distributions with a Gaussian noise $\mathcal{N}(\mu = (0,0), \Sigma = 0.001I_2)$. First, nine classical classifiers are trained and tested on these data. The standard 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 difficult to predict, i.e., most standard classifiers fail to predict the true class of the those samples.

The idea here for choosing the *ndc* hyper-parameter is to avoid misclassification when the samples are difficult. For the samples that are "certain", i.e., the posterior probability of one of the classes is close to 1, this later class obtain the maximum gain regardless the value given to β (see Subsection 3.1). Consequently, it is more interesting to set the value of β regarding the less "certain" samples. 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 each less "certain" sample of the training data set using leave-one-out technique and the second component is the mean of the second maximal probabilities p^2, and so on. Thus, $p^f = (p^1, p^2, ...)$. To determine the less "certain" sample a threshold is considered and when the maximal probability is lower than this threshold then the sample is considered less certain. in the illustration, this threshold is fixed to 0.99. The value of β is considered as the

boundary behind which if the sample is less certain than the mean probabilities of less "certain" samples, we should predict the subset of the two first classes with maximal probabilities. Thus,

$$\beta_{ndc} = \sqrt{\frac{p^1 - p^2}{p^2}}. \tag{14}$$

In Figure 3, we present the prediction when $\beta = 2.571$ is determined as in Equation (14). The samples that are considered as difficult to predict by the point prediction classifier are labelled by their number in the dataset. Only the samples number 140 and 82 are errors in the predictions of *ndc* and only three less (not labelled as difficult) difficult samples are predicted as imprecise. Note that, the example 140 is an exception as its probability is significantly above the one of the reference probability for the wrong class label. Concerning the difficult samples, ten samples are predicted as subsets of two classes containing the true class and one as the whole set. For the case of *eclair*, we consider binary classifications

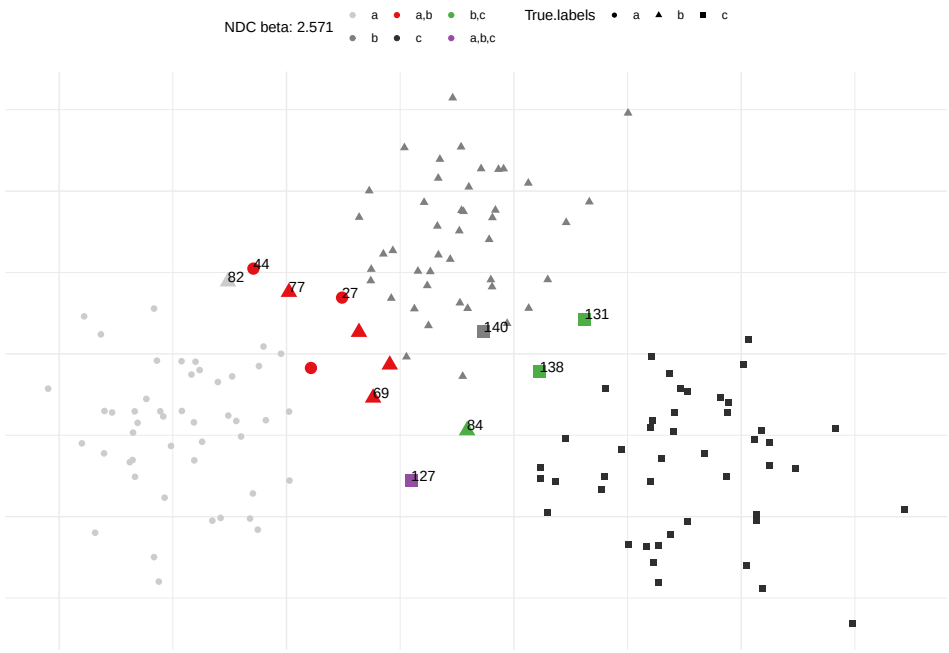

Fig. 3. The predictions obtained with *ndc*: a large size is given to the point symbols representing predictions that are errors or imprecise.

"a against b", "a against c" and "b against c". We apply the same reasoning by considering the leave-one-out technique to determine $m(\theta_1|x))$, $m(\theta_2|x))$ and

$m(\theta_{12}|x))$ for each example of the learning data set. Her also we consider only less certain samples with the same threshold. From the subsection 3.2, to avoid misclassification for difficult 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 less certain sample. The proposed value of β is β_{eclair} that is the solution the quadratic Equation (13) with $m(\theta_{12}|x)) = m_{12}$ and $m(\theta_1|x)) = 2\ (1 - m_{12})/3$. In Figure 4, we can see that, for the case "a against b", two predictions are still errors and four are imprecise. For the case of "a against c", we have only one imprecise prediction. While for the case "b against c", we have four imprecise predictions.

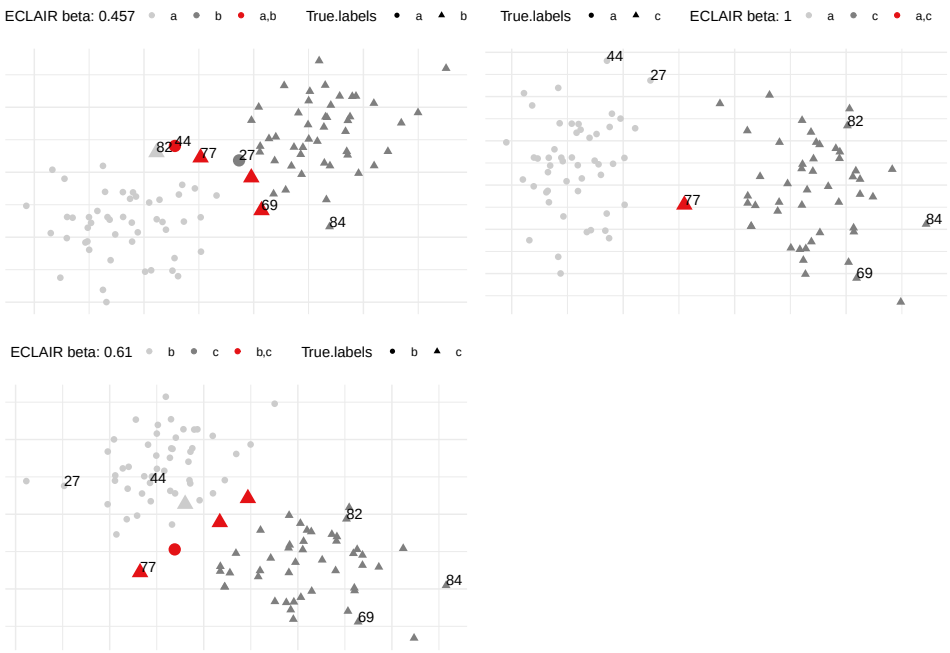

Fig. 4. The predictions obtained with *eclair*: a large size is given to the point symbols representing predictions that are errors or imprecise.

4.2 Illustration using UCI data

The second illustration concerns the comparison of the performances of the *ndc* classifier where β is determined as in the Equation (14) to the *ndc cv*, i.e., classifier tuned using cross-validation, and the naive credal classifier *ncc* using 11 UCI data based on the performances measures presented in the Subsection 2.4. The experimentation procedure is conduct as follows. Each dataset is split randomly 50 times to obtain a learning set (80%) and a testing set (20%). The

parameters are optimized, each time, using the cross-validation technique on the learning dataset. More precisely, for *ndc cv* two hyper-parameters are involved, the point prediction classifier used to obtain the posterior probabilities and the parameter β. For the first parameter, the choice is performed within the nine classifier presented in the Subsection 4.1 while for the second parameter the choice is performed in the interval $[0, 2]$ with steps of 0.1. Concerning *ncc*, the choice of parameter s is performed within a set of 20 values $S = \{10^{-30}, 10^{-20}, 10^{-15}, 10^{-10}, 10^{-9}, 10^{-8}, 10^{-7}, 10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 0.2, 0.3, 0.5, 0.6, 1, 1.1, 2\}$. The results are presented in Table 1. As one can see *ndc* gives the best result for the imprecise accuracy and u_{80} measure which means that it is more cautious than the two others while the its accuracies are still close to those of the best classifiers.

		Iris	BC	Wine	IS	DBT	Glass	PID	Sonar	Seeds	Forest	Ecoli
Accuracy	*ndc*	95.07	93.99	95.88	84.49	95.07	55.55	59.59	72.73	92.48	82.08	84.16
		± 3.76	± 2.43	± 3.66	± 3.72	± 4.5	± 7.43	± 9.13	± 7.63	± 3.38	± 4.42	± 6.74
	ndc cv	97	96.24	97.47	93.03	97.29	73.65	70.26	75.41	95.76	86.84	85.72
		± 3.03	± 1.55	± 2.14	± 2.99	± 3.27	± 7.75	± 6	± 7.26	± 2.6	± 2.96	± 3.96
	ncc	90.73	95.47	88.53	61.97	85.71	30.35	15.26	26.63	82.86	25.44	39.41
		± 4.92	± 2.26	± 5.52	± 9.14	± 7.7	± 20.71	± 4.06	± 11.57	± 5.81	± 6.75	± 12.28
$u50$	*ndc*	96.60	95.70	97.44	91.17	95.82	72.07	73.92	81.83	94.75	87.33	87.03
		± 2.82	± 1.47	± 2.18	± 2.03	± 3.28	± 4.61	± 3.07	± 4.77	± 2.79	± 2.34	± 3.98
	ndc cv	97.3	96.66	97.59	94.07	97.57	79.14	77.53	82.1	96.1	88.43	87.87
		± 2.79	± 1.34	± 2.05	± 2.03	± 2.77	± 4.81	± 3.18	± 4.85	± 2.57	± 2.34	± 3.36
	ncc	93.6	96.02	92.6	75.3	89.45	36.74	55.32	59.34	87.07	53	57.92
		± 3.47	± 1.63	± 3.84	± 5.14	± 5.11	± 13.24	± 1.96	± 2.17	± 4.24	± 2.51	± 6.74
$u65$	*ndc*	97.06	96.22	97.91	93.18	96.05	77.26	78.22	84.56	95.44	88.96	87.9
		± 2.66	± 1.32	± 1.84	± 1.7	± 3.06	± 4.53	± 3.17	± 4.72	± 2.78	± 2.01	± 3.45
	ndc cv	97.39	96.79	97.62	94.38	97.66	80.81	79.71	84.1	96.2	88.91	88.52
		± 2.77	± 1.34	± 2.05	± 1.89	± 2.66	± 4.74	± 3.18	± 4.87	± 2.6	± 2.31	3.38±
	ncc	94.51	96.19	94.07	79.3	90.65	39.65	67.34	69.15	88.38	61.83	63.9
		± 3.25	± 1.51	± 3.27	± 4.78	± 4.75	± 9.88	± 1.74	± 3.79	± 4.04	± 2.08	± 5.94
$u80$	*ndc*	97.52	96.73	98.38	95.18	96.27	82.45	82.52	87.29	96.12	90.58	88.77
		± 2.58	± 1.29	± 1.6	± 1.56	± 2.95	± 4.93	± 4.6	± 5.15	± 2.86	± 1.96	± 3.22
	ndc cv	97.48	96.91	97.66	94.7	97.74	82.48	81.89	86.11	96.3	89.39	89.16
		± 2.76	± 1.36	± 2.06	± 1.85	± 2.57	± 5.17	± 3.72	± 5.29	± 2.64	± 2.37	± 3.49
	ncc	95.43	96.36	95.54	83.3	91.84	42.56	79.36	78.97	89.7	70.66	69.87
		± 3.2	± 1.43	± 2.77	± 5.05	± 4.72	± 6.99	± 1.89	± 6.81	± 4.05	± 2.76	± 6.03
imprAcc	*ndc*	98.13	97.42	99	97.86	96.57	90.95	88.26	90.93	97.05	93.18	89.97
		± 2.62	±1.44	± 1.53	± 1.74	± 2.97	± 6.31	± 7.24	± 6.29	± 3.1	± 2.31	± 3.5
	ndc cv	97.6	97.08	97.71	95.11	97.86	84.85	84.79	88.78	96.43	90.04	90.03
		± 2.78	± 1.44	± 2.08	± 1.98	± 2.5	± 6.48	± 4.95	± 6.33	± 2.73	± 2.59	± 3.79
	ncc	97	96.58	99.18	88.63	93.93	60.85	95.38	92.05	91.76	87.09	81.16
		± 3.52	± 1.41	± 1.79	± 6.22	± 5.22	± 22.49	± 2.54	± 11.13	± 4.44	± 5.38	± 8.36

Table 1. The imprecise classifiers' performances on the UCI data.

5 Conclusion

In this paper we are interested in the imprecise classification. Especially, we focus on the study of the parameter β involved in the gain function used in the decision step of two imprecise classifiers. More precisely, we studied the predicted subsets depending on this parameter. We proposed a technique to choose the value of this parameter when the classifiers are involved in a classification task. Furthermore, the built classifiers give reasonable good performances related to evaluation measures for imprecise classifier.

References

1. Abellan, J., Masegosa, A.R.: Imprecise classification with credal decision trees. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems **20**(05), 763–787 (2012)
2. Couso, I., Sánchez, L.: Machine learning models, epistemic set-valued data and generalized loss functions: an encompassing approach. Information Sciences **358**, 129–150 (2016)
3. Coz, J.J.d., Díez, J., Bahamonde, A.: Learning nondeterministic classifiers. Journal of Machine Learning Research **10**(Oct), 2273–2293 (2009)
4. Imoussaten, A., Jacquin, L.: Cautious classification based on belief functions theory and imprecise relabelling. International Journal of Approximate Reasoning **142**, 130–146 (2022)
5. 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. (eds.) Scalable Uncertainty Management. pp. 122–135. Springer International Publishing, Cham (2019)
6. Jacquin, L., Imoussaten, A., Trousset, F., Perrin, D., Montmain, J.: Control of waste fragment sorting process based on mir imaging coupled with cautious classification. Resources, Conservation and Recycling **168**, 105258 (2021)
7. Ma, L., Denoeux, T.: Partial classification in the belief function framework. Knowledge-Based Systems p. 106742 (2021)
8. Quost, B., Masson, M.H., Destercke, S.: Dealing with atypical instances in evidential decision-making. In: International Conference on Scalable Uncertainty Management. pp. 217–225. Springer (2020)
9. Sanchez, L., Couso, I.: A framework for learning fuzzy rule-based models with epistemic set-valued data and generalized loss functions. International Journal of Approximate Reasoning **92**, 321–339 (2018)
10. Tsoumakas, G., Vlahavas, I.: Random k-labelsets: An ensemble method for multilabel classification. In: European conference on machine learning. pp. 406–417. Springer (2007)
11. Yang, G., Destercke, S., Masson, M.H.: The costs of indeterminacy: How to determine them? IEEE transactions on cybernetics **47**(12), 4316–4327 (2016)
12. Zaffalon, M.: A credal approach to naive classification. In: ISIPTA. vol. 99, pp. 405–414 (1999)
13. Zaffalon, M.: Statistical inference of the naive credal classifier. In: ISIPTA. vol. 1, pp. 384–393 (2001)
14. Zaffalon, M., Corani, G., Mauá, D.: Evaluating credal classifiers by utility-discounted predictive accuracy. International Journal of Approximate Reasoning **53**(8), 1282–1301 (2012)