Cautious classification based on belief functions theory and imprecise relabelling

Abdelhak Imoussaten, Lucie Jacquin

Cautious classification based on belief functions theory and imprecise relabelling

Abdelhak Imoussaten *, Lucie Jacquin

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

A B S T R A C T

The performances of standard classifiers, i.e., any method of point prediction for classifi-cation, decline in case of imperfect data. In some sensitive domains where these imper-fections are present, these classifiers need to be adapted in order to avoid any misclassi-fication that has serious consequences. Recent works proposed to deal with this problem by using cautious classification techniques. This paper is in line with these works, espe-cially with imprecise classifiers, i.e., the output of the classifier for an input sample that is subject to considerable imperfections is a subset of classes. The distinctive feature of our imprecise classification proposition is that it considers that in some applications, data im-perfection is not limited to new samples to be classified but can also be present in training data. We therefore propose a relabelling procedure which allows us to identify imperfect samples in the training data and relabel them with an appropriate subset of candidate classes. This approach to imprecise classification is close, in some aspects, to hierarchical classification where a "parent" can be considered as a subset of classes that are the "chil-dren" in the leaves. Furthermore, the belief functions framework is considered to represent the uncertainty and imprecision about the class of a new sample where the focal elements are contained in the set of new labels of the training data. A criterion based on a gener-alised F_β score and the obtained mass function is established to decide which subset of classes should be associated to the new sample. Several options are presented to build our classifier for the relabelling procedure and for the reasoning step. Thus, the performances of each option are presented before comparison with state-of-the-art imprecise classifiers' performances. The comparisons are conducted first on randomly generated data and then on 11 UCI datasets based on five measures of imprecise classification performances. They show that our classifier achieves performances close to, sometimes better than, the best on the five measures.

Keywords:
Imprecise classification
Cautious classification
Belief functions theory

1. Introduction

Several causes can lead to imperfect data such as measurement inaccuracy due to the quality of sensors, data unreliability, partial data, etc. In [1] the distinction is made when representing imperfect data between ontic and epistemic views [2]. In the ontic view, the imperfect data representation is interpreted as the "true value", i.e., a precise representation of reality while, in the epistemic view, the imperfect data representation is used to describe imperfection of knowledge

* Corresponding author.
E-mail addresses: abdelhak.imoussaten@mines-ales.fr (A. Imoussaten), lucie.jacquin@mines-ales.fr (L. Jacquin).

about the data. In this paper we consider a representation of data imperfection that is consistent with the epistemic view. In the cases where the data are imperfect, classical machine learning techniques reach their limits. Indeed, in the case of supervised classification, for example, a method of point prediction for classification can misclassify a new sample. This can have serious consequences when a classification task is involved in applications that are sensitive, as in medical diagnosis applications when the classifier has to detect early-stage cancer; or in autonomous car applications when the classifier has to distinguish between a human or animal and an object in dangerous situations; or in environmental compliance when a classifier has to sort plastics for recycling purposes, and identification errors will cause serious recycling difficulties and significant degradation of the secondary raw material performances [3], etc. The misclassification of some samples can be caused by the existence of *overlapping regions* in the feature space representation [4] [5] due to the imperfect data. This problem occurs when the obtained characteristics of some samples are very similar even if they are labelled using different class labels. The misclassification can also be caused when the samples belong to an *isolated region*, i.e., a region that is represented by very few samples in the training data.

Facing the issue of imperfect data, recent works focus on cautious classification. Cautious classification aims to minimize errors by providing a reliable output based on an appropriate representation of uncertainty. The idea of favouring the reliability of the information rather than the precision of its content was initially introduced in the work of R. A. Fisher [6] and J. Neyman [7] in the fields of statistical/Bayesian inference with the concept of confidence interval/region by associating a confidence threshold to the set provided as an estimation of a parameter. A particular case of cautious classification is imprecise classification, which consists in predicting a subset of candidate classes for a sample in the event of imperfect data. This can be beneficial, for example, in medical diagnosis applications by orienting the investigation toward a small set of candidate alternatives; or in the case of plastics sorting by adding a container dedicated to a given family of plastics [3], etc. The first works were proposed in the probability theory framework and consisted in predicting a subset of classes for a sample when the maximal probability is below a fixed threshold [8] [9] [10]. The predicted subset is the smallest whose cumulative probability exceeds a given threshold [10]. In a recent work [11], the latter approach was adapted to propose the *non-deterministic* classifier (ndc) by taking into account a gain function when maximising the expected gain. Based also on the framework of probability theory and statistics, conformal prediction [12] [13] provides confidence regions based on statistical hypothesis testing. More recently, the authors of [14] proposed, within the framework of belief functions, an extension of the utility matrix that considers the gains in the case of point prediction classification to an utility matrix considering gains in the case of imprecise prediction using aggregation functions. Within the framework of imprecise probabilities, imprecise classification is mainly based on a binary relation defined on the set of classes and the pre-order inferred from this relation. The subset of the non-dominated classes is considered as the prediction for the new sample. The *Naive Credal Classifier* (ncc) [15] [16] is an example of a classifier based on this approach where the dominance relation is inferred from the credal set representing the imprecision and uncertainty about the true class of a sample. The difficulty in this approach lies in the step of building the credal set. In [15] the credal set is built using the Imprecise Dirichlet Model (IDM) [17] while in [16] the credal set is built using an interval constraints approach [18]. Following the same principle, [19] proposes an imprecise classifier where the dominance relation is based on different quantifications of uncertainty based on a binary classifier [20] trained to distinguish aleatory and epistemic uncertainty. Furthermore, [21] proposed building a pre-order on the set of class labels based on a mass function. This approach uses interval dominance, where the intervals are represented by the belief and plausibility functions. The same principle of dominance is also used in [22] [23], where a fuzzy preference relation is deduced from the scores of a binary classifier in a "one-against-one" scheme. This fuzzy relation is exploited to obtain a pre-order on the set of classes.

Regarding what has been proposed in the state-of-the-art works, the distinguishing features of our proposition are as follows. First, we consider that imperfection can be present not only in the new samples but also in the training data. A relabelling procedure is therefore proposed to relabel the samples that are in *overlapping* or in *isolated* regions. Second, a gain matrix that incorporates the *specialisation* gain to control the trade-off between cautiousness and efficiency of the imprecise prediction is proposed. The second point is inspired both from the matrix gain proposed in [11] and the performance measures used in hierarchical classification. The *specialisation gain* is employed to penalise predictions that are more precise than the information available about a sample could allow. Moreover, the *relabelling step* allows us to directly learn labels in the form of subsets of classes, including singletons. This differentiates our approach from the other ones as only the subsets of classes identified in the relabelling step can be predicted to the new samples. The proposed approach is named *eclair* (Evidential CLAssification with Imprecise Relabelling) and uses the belief functions framework [24] to represent the uncertainty and imprecision present in the data. It is built in three steps: 1) *the relabelling step* which is applied to the learning data and serves to assign new labels, in the form of subsets of class labels, to the samples belonging to the overlapping and isolated regions in the feature space; 2) *the learning step* consists in training a standard classifier on the resulting new training data; 3) and the *reasoning step* which takes a mass function and a gain function as inputs and provides a subset of classes as output. When applying the reasoning step to process a new sample, the mass function corresponds to the chances quantified by the trained classifier for each subset of classes to be the "true" label of the sample. As one can notice in the mentioned works about imprecise classification, the classifiers fall in two categories. The dominance-relation based classifiers and the expected-gain based classifiers. The comparisons of *eclair* performances are presented related to two classifiers each representing one of the two categories: *ncc* for the first category and *ndc* for the second one.

This paper is organised as follows. In Section 2, some reminders are given about imprecise and hierarchical classification, elements of belief functions theory, and the evaluation measures of imprecise classifiers. In Section 3, the proposed approach

is detailed. Section 4 is dedicated to comparing firstly the different options of the *eclair* classifier, and secondly times to comparing *eclair* to the approaches of state-of-the-art imprecise classifiers. The comparison uses both a simulated dataset and UCI datasets. Finally a discussion is proposed in Section 3.4.

2. Background

In this section, background notions are briefly presented and notations are introduced. The imprecise classification is first presented in Section 2.1, in Section 2.2 the link between hierarchical classification and the imprecise classification is briefly discussed, some reminders about the theory of belief functions are given in Section 2.3 and the evaluation measures for the imprecise classifiers are presented in Section 2.4.

2.1. Imprecise classification

Imprecise classification is a kind of cautious classification that enables the prediction of a subset of candidate classes for *difficult* samples, i.e., samples for which the available information is highly affected by imperfections and make the assignment to a single class label too uncertain. More precisely, let us consider the case of a supervised classification task where the samples represented by P characteristics $X_1, X_2, \ldots, X_P$ should be classified using a set of n class labels $\Theta = \{\theta_1, \theta_2, \ldots, \theta_n\}$. Let us also consider a set of interest $\mathbb{A} \subseteq 2^\Theta \setminus \emptyset$. An *imprecise classifier* δ_{ic} is a mapping from the Cartesian product $\mathcal{X} = \prod_{i=1}^{P} X_i \subseteq \mathbb{R}^P$ to $\mathbb{A}$:

$$\delta_{ic} : \mathcal{X} \rightarrow \mathbb{A},$$

where for a sample $\boldsymbol{x} \in \mathcal{X}$, $\delta_{ic}(\boldsymbol{x}) \in \mathbb{A}$ is a subset of classes. Note that the number of elements in $\delta_{ic}(\boldsymbol{x})$ is not necessary strictly greater than one and unlike what is usually considered in the state of the art, we consider that the set of interest $\mathbb{A}$ is not necessary equal to $2^\Theta \setminus \emptyset$. Indeed, with regard to the interest of the user or the constraints of the application, the set $\mathbb{A}$ can be reduced to a very small part, necessarily including the singletons, of 2^Θ.

2.2. Hierarchical classification vs imprecise classification

In the hierarchical classification setting, the classes correspond to structured nodes governed by "IS-A" links called *parents-children* where the class *children* is included in the class *parents*. The nodes that have no children are called *leaves* of the hierarchical structure. Imprecise classification can be connected to hierarchical classification by considering that the prediction of a *parent* node in the hierarchy corresponds to the imprecise prediction of the subset containing all the *child* leaves attached to this node. Nevertheless, two main differences between these two classification techniques have to be mentioned. First, in the case of hierarchical classification, the predictions are elements of the pre-established hierarchical structure whereas for imprecise classification the outputs can be any non-empty subset of Θ, even if it has no particular semantics in reality. Second, according to [25], the case of partially labelled samples, i.e., whose true class label is not necessarily a leaf of the hierarchy, has been considered in several hierarchical classification works [26,27]. For this second point, the difference lies in the fact that in the hierarchical classification the nodes are governed by a conceptual representation [28] and a parent node could be the true class label for a sample in the training data. Consequently, two types of supervised classification errors are considered in hierarchical classification. The first one is the *generalisation error* [25] [29] that occurs when a classifier predicts a class that is the antecedent of the true class of a sample. The second one is the *specialization error* [25] [29] that occurs when a classifier predicts a class that is the descendant of the true class of a sample. Note that the term of *generalisation error* introduced in the hierarchical classification techniques should not be confused with the notion of the *generalisation error* that is central in machine learning and that quantifies the ability of an algorithm to predict the outcome of new samples. In the case of imprecise classification, the situation of partially labelled sample in training data occurs when this sample could not be precisely assigned to a single class label due to lack of knowledge. In this case, the *generalisation error* correspond to the error of predicting a subset of candidate class labels that is wider than the partial label of the sample. While the *specialization error* can be interpreted in the imprecise classification as the error of predicting a subset of candidate class labels that is smaller than the available information allows. Thus, when resolving an imprecise classification problem the situation can be similar to that of hierarchical classification in the case where the class labels of some samples in the training data are partially known.

2.3. Belief functions theory

In this Section we give a brief reminder about belief functions theory. Let us consider a reference set Θ where each element $\theta \in \Theta$ represents the lowest level of discernible information in Θ. The set Θ is called a frame of discernment. Belief functions theory, is an interesting framework to represent and process uncertain and imprecise information and can be seen as an extension of many uncertainty theories. The recent growing interest in this theory has allowed techniques to be developed to resolve a diverse range of problems such as estimation [30] [31][32], standard classification [33,34], or even

hierarchical classification [35,36]. Three main set functions are involved in the belief functions framework. The *mass function* which assigns probabilities to imprecise information, leading to the distinction between equiprobability and imprecision or ignorance. More precisely, a *mass function*, also called *basic belief assignment* (bba), is a set function $m : 2^{\Theta} \rightarrow [0, 1]$ satisfying

$$m(\emptyset) = 0 \text{ and } \sum_{A \subseteq \Theta} m(A) = 1.$$

The elements $A \in 2^{\Theta}$, such that $m(A) > 0$, are called focal elements and they form a set denoted $\mathbb{F}$. The pair $(m, \mathbb{F})$ is called the body of evidence. The *belief function* measures the quantity of evidence proving an event, $Bel : 2^{\Theta} \rightarrow [0, 1]$, satisfying

$$Bel(A) = \sum_{B \subseteq \Theta, B \subseteq A} m(B).$$

The *plausibility function* measures the quantity of evidence that makes an event possible, $Pl : 2^{\Theta} \rightarrow [0, 1]$, satisfying

$$Pl(A) = \sum_{B \subseteq \Theta, B \cap A \neq \emptyset} m(B).$$

The above-defined set functions constitute the credal level where beliefs are captured and quantified. A second level considered in the belief functions framework is the pignistic level or decision level where beliefs are quantified using probability distributions. In this latter level, the pignistic probability distribution is computed as follows: $\forall \theta \in \Theta$,

$$pig_m(\theta) = \sum_{A \subseteq \Theta, A \ni \theta} \frac{m(A)}{|A|},$$

where $|A|$ denotes the number of elements in A.

2.4. Evaluation measures for the imprecise classifiers

When a classifier provides an imprecise prediction for a new sample, the evaluation of its performance is not straightforward and the classical measure of *accuracy* is not appropriate. The task of evaluating the performance of an imprecise classifier can be formulated as the problem of defining a performance measure that takes into account the trade-off between the criterion of cautiousness, i.e., the classifier should predict a subset of classes including the "true" classes of a sample that has imperfect data, and the criterion of efficiency, i.e., the predicted subset of classes should be as small as possible depending on the sample data. Several works have studied this problem and provide some measures to model this trade-off [16], [37], [38]. *Discounted accuracy* [39] seems to be an interesting measure as it takes into account the size of the predicted subset. However, as pointed out in [40], this measure rewards a predicted subset as if a random assignment had been made in this subset. Thus the discounted accuracy fails to recognize the benefit of cautious decisions over hazardous decisions. To handle this problem new measures were proposed to increase the cautiousness reward to the necessary extent, i.e., the degree to which the decision maker prefers to fix the reward of cautiousness depending on his application and the quality of the information obtained for the samples. These measures are represented by a function g of the *discounted accuracy* 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 [40] 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 formula of the above-mentioned performance evaluation measures applied to classifier δ_{ic} and the test data dst are given as follows:

- the *classical accuracy* measure that quantifies the proportion of good predictions:

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

- the *imprecise accuracy* (imprAcc) measure that quantifies the predictions containing the true class labels of the test samples:

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

- the *discounted accuracy* (discAcc) measure that quantifies the mean of the proportions of the good predictions in the predicted subsets and corresponds to the function $g(z) = z$:

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

This measure is denoted also u_{50}.

- The u_{65} measure that corresponds to the function $g(z) = -0.6 \cdot z^2 + 1.6 \cdot z$:

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

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

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

Where for a subset A of Θ, the function $\mathbb{1}_A : \Theta \to \{0, 1\}$ is the characteristic function such that: $\mathbb{1}_A(\theta) = 1$ if $\theta \in A$; and 0 otherwise. The u_{50} measure is considered as well suited to binary classification problems while u_{65} and u_{80} are well suited to classification problems with more than two class labels [40].

3. The *eclair* classifier

As it is briefly presented in Section 1, the *eclair* approach is built in three steps: 1) *the relabelling step*, 2) *the learning step*, 3) and the *reasoning step*. This Section presents these three steps. More precisely, let us consider a data set of L learning samples $(\boldsymbol{x}^l, \theta^l)_{1\leq l\leq L}$ where for each $l, 1 \leq l \leq L$, $\boldsymbol{x}^l \in \mathcal{X}$ and $\theta^l \in \Theta$. The *relabelling* of a learning sample $\boldsymbol{x}^l$ consists of assigning to it the smallest subsets of classes expressing the imprecision of its characteristics if it belongs to the overlapping or the isolated regions in the training data. We assume that the new subset of classes, denoted A^l, associated to the sample $\boldsymbol{x}^l$ necessarily contains the original class label θ^l.

Once the relabelling step is performed, we obtain a new learning samples $(\boldsymbol{x}^l, A^l)_{1\leq l\leq L}$ where for each $l, 1 \leq l \leq L$, $\boldsymbol{x}^l \in \mathcal{X}$ and $A^l \in \mathbb{A}$. The set of interest $\mathbb{A} \subseteq 2^\Theta \setminus \emptyset$ mentioned in the Subsection 2.1 is constituted by the distinct labels A^l of the training data set. The *learning step* then consists in training a learning algorithm of point prediction for classification to recognize the labels in $\mathbb{A}$. Finally, using the *reasoning step* we assign to a new sample $\boldsymbol{x}$ an element in $\mathbb{A}$ based on 1) the posterior mass function $m(.|\boldsymbol{x})$; 2) and an appropriate gain matrix representing the trade-off between cautiousness and efficiency. Note that the mass function $m(.|\boldsymbol{x})$ is obtained from the posterior probability mass function $p(.|\boldsymbol{x}) : \mathbb{A} \to [0, 1]$ provided by a method of point prediction quantifying the chance of each element from $\mathbb{A}$ to be the "true" label of $\boldsymbol{x}$. The focal elements of $m(.|\boldsymbol{x})$ are necessarily elements in $\mathbb{A}$. Note also that the idea of using a method of point prediction in an imprecise classification task is also used in [11] and [12]. In the following subsections the relabelling, the training and the reasoning steps are presented in detail.

3.1. Relabelling step

The relabelling procedure is usually used to identify suspect samples with the intention to remove or relabel them into a concurrent, more appropriate class label [41] [42] [43]. Closer to our work are the relabelling procedures proposed in [44] and [45] that are taking advantage of the information provided by the k nearest neighbours (knn) of the treated sample. In [45] the sample is relabelled using the majority vote of the k nearest neighbours but this technique may not be effective when some neighbours of the sample are non-representative of their classes or when the sample belongs to an isolated region. In [44] an evidential relabelling is performed on the training data using the evidential k nearest neighbours (EKNN) [33]. This evidential relabelling is not suitable for the implementation of the two other steps of *eclair* classifier where the partially labelled training samples are used to learn a point prediction for classification method.

Two new relabelling procedures are proposed in this paper as a generalisation of the one based on the knn of the sample. Both procedures are based on the posterior probability distribution provided by a method of point prediction for classification and the cross-validation technique. Indeed, knowing that the posterior probability distribution summarizes the information about the chances of each class to be the "true" class of a sample, one can use it to identify the subsets of classes that are candidates to be the new label of the sample.

Let us consider a sample $\boldsymbol{x}^l$ from the learning dataset and a method of point prediction δ_c. Using the leave-one-out technique, $\boldsymbol{x}^l$ is removed from the learning data and its predicted class is $\delta_c(\boldsymbol{x}^l)$. We denote by $p_c(\theta_i|\boldsymbol{x}^l)$ the posterior probability quantifying the chances of each class label θ_i, $i \in \{1, \ldots, n\}$, to be the "true" class of $\boldsymbol{x}^l$. Let us denote by A^l the new label of $\boldsymbol{x}^l$. The first method, called the *rank method*, consists in including in A^l the classes having a posterior probability at least equal to the one obtained by the true class θ^l:

$$A^l = \{\theta \in \Theta, p_c(\theta|\boldsymbol{x}^l) \geq p_c(\theta^l|\boldsymbol{x}^l)\}. \tag{1}$$

Obviously, $\delta_c(\boldsymbol{x}^l)$ and θ^l are in A^l. In the most of cases, i.e., when data are perfect, $A^l = \{\delta_c(\boldsymbol{x}^l)\} = \{\theta^l\}$. The second method, called the *entropy method*, is based on the Shannon entropy index and is detailed below. The Shannon entropy, also called *entropy index* denoted H, quantifies for a probability distribution the degree of perfection with which an outcome is predicted. If $H(p_c(.|\boldsymbol{x}^l)) = -\sum_{i=1}^{n} p_c(\theta_i|\boldsymbol{x}^l)\log(p_c(\theta_i|\boldsymbol{x}^l))$ (only coefficients with $p_c(\theta_i|\boldsymbol{x}^l) > 0$ are considered) is close to 0 then

Table 1
The relabelling of some samples using the two methods.

sample	class	posterior probability				H	relabelling	
		θ_1	θ_2	θ_3	θ_4		*rank method*	*entropy method*
$\boldsymbol{x}^1$	θ_4	0.2	0.25	0.25	0.3	1.376	$\{\theta_4\}$	Θ
$\boldsymbol{x}^2$	θ_3	0	0	0.99	0.01	0.056	$\{\theta_3\}$	$\{\theta_3\}$
$\boldsymbol{x}^3$	θ_1	0.1	0	0.9	0	0.325	$\{\theta_1, \theta_3\}$	Θ
$\boldsymbol{x}^4$	θ_1	0.2	0	0.8	0	0.5004	$\{\theta_1, \theta_3\}$	$\{\theta_1, \theta_3\}$
$\boldsymbol{x}^5$	θ_2	0.2	0.25	0.5	0.05	1.165	$\{\theta_2, \theta_3\}$	$\{\theta_1, \theta_2, \theta_3\}$

the prediction from $p_c(.|\boldsymbol{x}^l)$ is perfect while if $H(p_c(.|\boldsymbol{x}^l))$ is high the uncertainty about the predicted class for $\boldsymbol{x}^l$ is also too high. The proposition of relabelling $\boldsymbol{x}^l$ consists in fixing a threshold ρ of the *entropy index* beyond which the degree of uncertainty is considered too high. When the entropy is greater than ρ, the predicted class of $\boldsymbol{x}^l$ is questionable. To simplify the notations, we introduce the following quantity related to the *entropy index* for a subset $B \subset \Theta$:

$$H_B(p_c(.|\boldsymbol{x}^l)) = -\mathbb{P}_c(B|\boldsymbol{x}^l)\log(\mathbb{P}_c(B|\boldsymbol{x}^l)) - \sum_{\theta_i \in \Theta \setminus B} p_c(\theta_i|\boldsymbol{x}^l)\log(p_c(\theta_i|\boldsymbol{x}^l)), \tag{2}$$

where $\mathbb{P}_c(.|\boldsymbol{x}^l)$ is the probability measure associated to the probability distribution $p_c(.|\boldsymbol{x}^l)$. The Equation (2) can be seen as the *entropy index* of a new probability distribution deduced from $p_c(.|\boldsymbol{x}^l)$ where all the classes in B are considered as the same single class. Thus, the choice of the new label A^l of $\boldsymbol{x}^l$ depends on ρ, $\delta_c(\boldsymbol{x}^l)$ and $p_c(.|\boldsymbol{x}^l)$. According to whether $\delta_c(\boldsymbol{x}^l)$ is equal to θ^l or not, two thresholds can be considered and respectively denoted ρ_1 and ρ_2. In the case where $\delta_c(\boldsymbol{x}^l)$ is θ^l, the proposed new label A^l is:

$$A^l = \begin{cases} A_{\rho_1}, & \text{if } H(p_c(.|\boldsymbol{x}^l)) > \rho_1, \\ \{\theta^l\} & \text{elsewhere.} \end{cases} \tag{3}$$

In the case where $\delta_c(\boldsymbol{x}^l)$ is different from θ^l, the proposed new label A^l is:

$$A^l = \begin{cases} \{\theta^l\} \cup A_{\rho_2}, & \text{if } H(p_c(.|\boldsymbol{x}^l)) > \rho_2, \\ \Theta & \text{elsewhere,} \end{cases} \tag{4}$$

where

$$A_\rho = \text{argmin}_{B \subset \Theta}\{|B|, H_B(p_c(.|\boldsymbol{x}^l)) \leq \rho\},$$

is the smallest subset for which if all its classes are considered as the same, the new entropy index, H_{A_ρ}, is below the threshold ρ for $p_c(.|\boldsymbol{x}^l)$.

In the case of the relabelling defined in Equation (3), the new label A^l is composed initially by θ^l and other classes are added until the new entropy $H_{A_{\rho_1}}$ becomes lower than ρ_1 and if A^l contains more than one class, $\boldsymbol{x}^l$ is considered as non-representative of the class θ^l. While, in the case of the relabelling defined in Equation (4), $\boldsymbol{x}^l$ is considered as non-representative of the class θ^l. If the entropy index is higher than ρ_2, then $\boldsymbol{x}^l$ is relabelled as the union of $\{\theta^l\}$ and A_{ρ_2} (A_{ρ_2} can contain θ^l). Otherwise, $\boldsymbol{x}^l$ is considered as too ambiguous and is relabelled by Θ.

Example 3.1. Let us consider five training samples and the corresponding posterior probability distribution presented in Table 1. These samples are part from the training samples $(\boldsymbol{x}^l, \theta^l)_{1 \leq l \leq L}$ and $|\Theta| = 4$. Note that the entropy method is performed using the parameters $\rho_1 = \rho_2 = 0.5$.

More generally, from Equation (3) the more the entropy threshold ρ_1 is very small, i.e., close to 0, the more the samples are relabelled with subsets of classes and from Equation (4), the more the entropy threshold ρ_2 is high, the more the samples are relabelled with Θ. Consequently, if a significant quantity of uncertain and ambiguous information is present in the training data, a large number of the samples are relabelled with subsets of classes. The choice of A^l in Equations (3) and (4) is governed by the assumption considered in this work that the original class θ^l of the sample is provided by a reliable source. Thus, this information is considered certain. Whereas, the measured characteristics of the sample are considered as potentially imperfect. It is certain that this choice leads to a loss of information in the training data mainly for the isolated samples and the samples belonging to overlapping regions, i.e., regions corresponding to boundaries between the classes. More precisely, adding another class to the original class θ^l when constituting the new label A^l impoverishes the original information. Nevertheless, as mentioned in the introduction of this paper, the proposition of this paper is designed to tackle decision problems involving sensitive applications where cautiousness is privileged. In such an application we assume that one prefers to be cautious than to bet on a class when the information is imperfect. A second assumption is made to simplify the approach which consists in also considering that the new label A^l is provided by a reliable source, i.e., the

relabelling procedure. However, a more general approach could cover the cases of non-reliable sources for the new labels. Such an approach could consist in assigning to each learning sample a mass function taking into account the reliability of the relabelling method.

3.2. Training step

Let us consider a learning samples $(\boldsymbol{x}^l, A^l)_{1\leq l\leq L}$ where for each $l, 1 \leq l \leq L$, $\boldsymbol{x}^l \in \mathcal{X}$ and $A^l \in \mathbb{A}$ where $\mathbb{A}$ is a subset of $2^\Theta \setminus \emptyset$ that necessary includes all the singletons, i.e., for each $\theta \in \Theta$, $\{\theta\} \in \mathbb{A}$. Let us consider a point prediction for classification method δ_c. Each label $A^l \in \mathbb{A}$ is a subset of Θ, so some labels can have intersections. When training δ_c on the data $(\boldsymbol{x}^l, A^l)_{1\leq l\leq L}$, the method considers that the elements of $\mathbb{A}$ are independent and the intersections of its elements are ignored in this step. Consequently, when dealing with a new sample $\boldsymbol{x}$, δ_c provides a posterior probability distribution $p(.|\boldsymbol{x}) : \mathbb{A} \to [0, 1]$ quantifying the chances of each element from $\mathbb{A}$ to be the "true" label of $\boldsymbol{x}$. When $\boldsymbol{x}$ is classified as a non-singleton label A, this means that the true class of $\boldsymbol{x}$ belongs to A and that the available information of the training data and the characteristics observed on $\boldsymbol{x}$ do not allow more precise classification for $\boldsymbol{x}$. As $p(.|\boldsymbol{x})$ quantifies the chances of subsets of Θ to contain the true class of $\boldsymbol{x}$, it can be considered as a mass function $m(.|\boldsymbol{x}) : 2^\Theta \to [0, 1]$ whose focal elements are included in $\mathbb{A}$. Note that all the point prediction for classification methods are adapted to provide a posterior probability distribution when dealing with a new sample to classify. Furthermore, this step could as well be performed using the evidential $k-NN$ rule for partially supervised data [46][47].

3.3. The reasoning step

This Section presents the reasoning step used in the *eclair* approach. It consists in associating a class or a subset of classes to a new sample $\boldsymbol{x}$ based on 1) a posterior mass function $m(.|\boldsymbol{x}) : 2^\Theta \to [0, 1]$ where the set of focal elements $\mathbb{F}(.|\boldsymbol{x})$ is such that $\mathbb{F}(.|\boldsymbol{x}) \subseteq \mathbb{A} \subseteq 2^\Theta$; and 2) a gain matrix $g : \mathbb{A} \times \mathbb{A} \to \mathbb{R}^+$ that associates to each pair $(T_{\boldsymbol{x}}, B)$ of two elements in $\mathbb{A}$ the obtained gain when the prediction is A such that the "true" label of $\boldsymbol{x}$ is $T_{\boldsymbol{x}}$. Three methods are proposed to perform the reasoning step in our approach: 1) the *01 gain* method which consists in predicting for $\boldsymbol{x}$ the subsets $A \in \mathbb{A}$ having the maximum mass function coefficient; 2) the *pignistic ndc* method which is similar to the decision step of *ndc* where the posterior probability distribution is replaced by the pignistic probability distribution obtained from $m(.|\boldsymbol{x})$ (see Subsection 3.3.1); 3) the generalised ndc-based F_β-score method that is presented in the Subsection 3.3.2. In the remainder of this section, the construction of the gain matrix in the case of the second and the third methods are presented. The gain matrix of the first method is obvious and is defined for $(T_{\boldsymbol{x}}, B) \in \mathbb{A} \times \mathbb{A}$ as follows:

$$g(T_{\boldsymbol{x}}, B) = g_{01}(T_{\boldsymbol{x}}, B) = \begin{cases} 1, & \text{if } B = T_{\boldsymbol{x}}, \\ 0 & \text{elsewhere.} \end{cases} \tag{5}$$

Finally, in the Subsection 3.3.3, the objective function used as the criterion to choose the optimal subsets in $\mathbb{A}$ as the prediction for $\boldsymbol{x}$ is presented for the three methods.

3.3.1. ndc based F_β-score

The gain matrix proposed in [11] is based on the F_β measure which is function of the *recall* and *precision* measures introduced in the domain of information retrieval. More precisely, the gain obtained when predicting $A \in \mathbb{A}$ such that the true class label is $\theta \in \Theta$ ($\theta \in A$) is given as follows:

$$F_\beta(\theta, A) = \frac{(1+\beta^2) \cdot recall(\theta, A) \cdot precision(\theta, A)}{\beta^2 \cdot precision(\theta, A) + recall(\theta, A)}, \tag{6}$$

where the *recall* is defined as the proportion of relevant classes included in A and the *precision* is defined as the proportion of retrieved classes in A that are relevant. In the case of imprecise classification these measures are given as: $recall(\theta, A) = \mathbb{1}_A(\theta)$ the proportion, i.e. 0% or 100%, of relevant classes that are included in A and $precision(\theta, A) = \frac{\mathbb{1}_A(\theta)}{|A|}$ the proportion of retrieved classes in A that are relevant which correspond respectively to the *discounted accuracy* and *imprecise accuracy* defined in the Subsection 2.4. Thus, we have:

$$F_\beta(\theta, A) = \frac{(1+\beta^2) \cdot \mathbb{1}_A(\theta)}{\beta^2 + |A|}. \tag{7}$$

The gain matrix defined by the F_β measure in Equation (7) is a trade-off between the *imprecise accuracy*, i.e., measure of cautiousness, and the *discounted accuracy*, i.e., measure of efficiency, where the parameter β is used to control the required levels of relevance and cautiousness.

Table 2
The quantification of the gain matrix using the function $F_{g,\beta}$.

		The predicted subset A						
		$\{\theta_1\}$	$\{\theta_2\}$	$\{\theta_3\}$	$\{\theta_1,\theta_2\}$	$\{\theta_1,\theta_3\}$	$\{\theta_2,\theta_3\}$	Θ
$T_{\boldsymbol{x}}$	$\{\theta_1\}$	1	0	0	$\frac{1+\beta^2}{2+\beta^2}$	$\frac{1+\beta^2}{2+\beta^2}$	0	$\frac{1+\beta^2}{3+\beta^2}$
	$\{\theta_2\}$	0	1	0	$\frac{1+\beta^2}{2+\beta^2}$	0	$\frac{1+\beta^2}{2+\beta^2}$	$\frac{1+\beta^2}{3+\beta^2}$
	$\{\theta_3\}$	0	0	1	0	$\frac{1+\beta^2}{2+\beta^2}$	$\frac{1+\beta^2}{2+\beta^2}$	$\frac{1+\beta^2}{3+\beta^2}$
	$\{\theta_1,\theta_2\}$	$\frac{1+\beta^2}{1+2\beta^2}$	$\frac{1+\beta^2}{1+2\beta^2}$	0	1	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1+\beta^2}{3/2+\beta^2}$
	$\{\theta_1,\theta_3\}$	$\frac{1+\beta^2}{1+2\beta^2}$	0	$\frac{1+\beta^2}{1+2\beta^2}$	$\frac{1}{2}$	1	$\frac{1}{2}$	$\frac{1+\beta^2}{3/2+\beta^2}$
	$\{\theta_2,\theta_3\}$	0	$\frac{1+\beta^2}{1+2\beta^2}$	$\frac{1+\beta^2}{1+2\beta^2}$	$\frac{1}{2}$	$\frac{1}{2}$	1	$\frac{1+\beta^2}{3/2+\beta^2}$
	Θ	$\frac{1+\beta^2}{1+3\beta^2}$	$\frac{1+\beta^2}{1+3\beta^2}$	$\frac{1+\beta^2}{1+3\beta^2}$	$\frac{1+\beta^2}{1+3/2\beta^2}$	$\frac{1+\beta^2}{1+3/2\beta^2}$	$\frac{1+\beta^2}{1+3/2\beta^2}$	1

3.3.2. The generalised F_β-score

The gain matrix defined in Equation (7) is based on the definition of the gain obtained in the case where only the chances of single class to be the true label of $\boldsymbol{x}$ are quantified. Considering that the chances of the element of $\mathbb{A}$ to be the "true" label of $\boldsymbol{x}$ are quantified by the mean of the mass function $m(.|\boldsymbol{x})$, the Equation (7) defining the gain matrix needs to be extended. A subset $A \in \mathbb{A}$ is considered as the "true" label of $\boldsymbol{x}$ means that the true class of $\boldsymbol{x}$ belongs to A. More precisely, when the true class is known precisely, we check if the information "$\theta_{\boldsymbol{x}} = \theta$" is true or false, where $\theta_{\boldsymbol{x}}$ is the unknown true class of $\boldsymbol{x}$ and $\theta \in \Theta$. In the case of imprecise information A, the information to check is "$\theta_{\boldsymbol{x}} \in A$". In the approach proposed in this paper, we take advantage of the specialisation and generalisation errors proposed in the hierarchical classification [25] to build the gain matrix. The *generalisation gain* is already exploited in Equation (7), i.e., rewarding the predictions that are efficient. We consider the same gain and we enrich it using the *specialisation gain*, i.e., rewarding the predictions that are cautious or penalize random choices in the predicted subsets of classes when the available information about the true class of $\boldsymbol{x}$ is imprecise.

Let us consider that given the available information, the appropriate but unknown label to associate to the sample $\boldsymbol{x}$ is $T_{\boldsymbol{x}} \in \mathbb{A}$. Each prediction $A \in \mathbb{A}$ such that $A \subset T_{\boldsymbol{x}}$ should be penalized in the *specialisation gain* and each prediction $B \in \mathbb{A}$ such that $B \supseteq T_{\boldsymbol{x}}$ should be penalized in the *generalisation gain*. The new gain matrix, denoted $F_{g,\beta}$, $F_{g,\beta} : \mathbb{A} \times \mathbb{A} \to \mathbb{R}^+$ is then defined for a "truth" $T_{\boldsymbol{x}} \in \mathbb{A}$, a prediction $A \in \mathbb{A}$ and a parameter $\beta \geq 0$ as follows:

$$F_{g,\beta}(T_{\boldsymbol{x}}, A) = \frac{(1+\beta^2)|A \cap T_{\boldsymbol{x}}|}{\beta^2|T_{\boldsymbol{x}}| + |A|}. \tag{8}$$

In this case the definitions of the measures of recall and imprecision are extended as follows: $recall(T_{\boldsymbol{x}}, A) = \frac{|A \cap T_{\boldsymbol{x}}|}{|T_{\boldsymbol{x}}|}$ and $precision(T_{\boldsymbol{x}}, A) = \frac{|A \cap T_{\boldsymbol{x}}|}{|A|}$. In other words, $recall(T_{\boldsymbol{x}}, A)$ is the proportion of relevant class labels $T_{\boldsymbol{x}}$ that are predicted and $precision(T_{\boldsymbol{x}}, A)$ the proportion of the predicted class labels A that are relevant.

Example 3.2. In the following an example, in the case of three classes $\Theta = \{\theta_1, \theta_2, \theta_3\}$, is given to show the quantification of the *gain matrix* especially to highlight the impact of *specialisation* and *generalisation* gains in Equation (8). As shown in Table 2, the first extreme case corresponds to the worst case that occurs when the "truth" $T_{\boldsymbol{x}}$ does not intersect with the predicted subset A, i.e., $T_{\boldsymbol{x}} \cap A = \emptyset$. In such a case the gain is minimal, i.e., $F_{g,\beta}(T_{\boldsymbol{x}}, A) = 0$ regardless the value of β. The second extreme case occurs when $T_{\boldsymbol{x}} = A$ and the gain is quantified as maximal $F_{g,\beta}(T_{\boldsymbol{x}}, A) = 1$ regardless the value of β. Concerning the intermediate cases, the generalisation gain guarantee the same results as in Equation (7), i.e., $F_{g,\beta}(T_{\boldsymbol{x}}, A) = \frac{(1+\beta^2)\ \mathbb{1}_A(T_{\boldsymbol{x}})}{\beta^2+|A|}$ ($T_{\boldsymbol{x}}$ is a single class). The new gains capturing the *specialization gain* can be seen in the cases where $A \subset T_{\boldsymbol{x}}$. For instance, if $A = \{\theta_1\}$ and $T_{\boldsymbol{x}} = \{\theta_1, \theta_2\}$, the associated gain is $F_{g,\beta}(\{\theta_1, \theta_2\}, \{\theta_1\}) = \frac{1+\beta^2}{1+2\beta^2}$ and in the case where $T_{\boldsymbol{x}} = \Theta$, the gain is $F_{g,\beta}(\Theta, \{\theta_1\}) = \frac{1+\beta^2}{1+3\ \beta^2}$. Thus, the larger the "truth" containing the prediction the smaller the gain function rewarding the good predictions. Also, with the new gain matrix, predictions containing more than one element can receive the maximal gain 1 in the case where they are equal to the "truth" $T_{\boldsymbol{x}}$ which is not possible when restricting the "truth" to the single classes. More generally, the value of β can be used to control the cautiousness and the relevance:

- when β is small, the prediction that is not efficient is penalised. Especially for $\beta = 0$, F_β boils down to precision.
- when β is large, the prediction that is not cautious is penalised. Especially for $\beta \to \infty$, F_β tends towards recall.

3.3.3. The objective function

Finally, the objective function used as the criterion to choose the optimal subsets in $\mathbb{A}$ as the prediction for $\boldsymbol{x}$ is the expected gain function based on the gain matrix defined for each method and the posterior mass function. For the *01 gain* method, the expected gain $EG(A|\boldsymbol{x})$ when choosing the imprecise prediction $A \in \mathbb{A}$ for $\boldsymbol{x}$ is:

$$EG(A|\boldsymbol{x}) := EG_{01}(A|\boldsymbol{x}) = \sum_{T_{\boldsymbol{x}} \in \mathbb{A}} m(T_{\boldsymbol{x}}|\boldsymbol{x}) \cdot g_{01}(T_{\boldsymbol{x}}, A) = m(T_{\boldsymbol{x}}|\boldsymbol{x}), \tag{9}$$

where the gain matrix g_{01} is defined in Equation (5). In the case of the *pignistic ndc* method, the expected gain is defined as follows for $A \in \mathbb{A}$:

$$EG(A|\boldsymbol{x}) := EG_{\beta}(A|\boldsymbol{x}) = \sum_{\theta_{\boldsymbol{x}} \in \Theta} pig_m(T_{\boldsymbol{x}}|\boldsymbol{x}) \cdot F_{\beta}(\theta_{\boldsymbol{x}}, A), \tag{10}$$

where the gain matrix F_β is defined in Equation (7). The expected gain for the generalised F_β-score based method is a generalisation of the expected gain defined in Equation (9) and is defined as follows for $A \in \mathbb{A}$:

$$EG(A|\boldsymbol{x}) := EG_{g\beta}(A|\boldsymbol{x}) = \sum_{T_{\boldsymbol{x}} \in \mathbb{A}} m(T_{\boldsymbol{x}}|\boldsymbol{x}) \cdot F_{g,\beta}(T_{\boldsymbol{x}}, A), \tag{11}$$

where the gain matrix $F_{g,\beta}$ is defined in Equation (8). Finally, for the three methods, the *eclair* imprecise prediction $\delta_{\beta}^{eclair}(\boldsymbol{x})$ for the sample $\boldsymbol{x}$ is given as:

$$\delta_{\beta}^{eclair}(\boldsymbol{x}) = \text{argmax}_{A \in \mathbb{A}}\, EG(A|\boldsymbol{x}). \tag{12}$$

3.4. Discussion

The proposed classifiers based on the *eclair* approach are comparable to the approaches of the state of the art on some aspects. On the one hand, they are close to the approaches of the hierarchical classification related to the idea of the imprecise or partial labelling of the training data and close to the approaches based on the imprecise probabilities concerning the representation of imprecision [15]. On the other hand, they are close to those using a gain matrix based on the F_β score to model the trade-off between cautiousness and efficiency [11]. However, the computational complexity of the *eclair* classifiers and *ncc* classifier, as for any approach representing imprecision in the data, can be very high. Indeed, if $n = |\Theta|$ is very large and $|\mathbb{A}| \gg n$, the computational complexity of the reasoning step of the *eclair* classifiers becomes very high, i.e., $O(|\mathbb{A}|^2)$, at worst exponential and at best quadratic. The computational complexity is lower when the gain function is 01, i.e., $O(|\mathbb{A}|)$, or when the gain function is based on the pignistic probability distribution, i.e., $O(n \cdot |\mathbb{A}|)$. These two last types of classifiers can be preferred when n is very large. The *ndc* classifier has the lowest computational complexity, i.e., at worst $O(n)$ [11], and the *ncc* one is quadratic, i.e., $O(n^2 - n)$ [15].

Furthermore, the imprecise predictions provided by the *eclair* classifiers are based on the maximisation of the expected gain where the chances are quantified by the posterior mass function and the matrix gain is built using the F_β score. For other choices, some criteria can be found in [14] that could be used to decide on the imprecise classification when the information concerning the possible states of nature is presented by a mass function.

4. Illustration and comparisons

In this illustration Section we present two parts. The first part is dedicated to analysing the imprecise classifiers' results on a simulated data and the second part presents the comparisons on the UCI data benchmark. Table 3 summarises some information about the methods of point prediction for classification and the imprecise classifiers that are involved in this section. Note that all the implementations are performed using R packages with the default parameters for the methods of point prediction. Note also that for the illustrations where the parameters ρ_1 and ρ_2 involved in the entropy based relabelling method are considered equal, they are both denoted ρ. Furthermore, in the studied data, some new labels obtained from the relabelling step could be less represented in the learning data set so we are faced with an *imbalanced data* problem. To deal with this problem, the Synthetic Minority Over-sampling Technique (SMOTE) [48], which consists in creating synthetic examples to increase the representativeness of minority class labels, is used.

4.1. Illustration using simulated data

In this first illustration, a simulated data for three classes a, b, and c is considered. For each class 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 a, $\mathcal{N}(\mu_b = (0.5, 0.9), \Sigma_b = 0.01I_2)$ for the class b, and $\mathcal{N}(\mu_c = (0.8, 0.6), \Sigma_c = 0.01I_2)$ for the class c. First, the results of the relabelling procedure are presented. Table 4 and Fig. 1 give the resulting new labels for four methods: method rank, method entropy with $\rho = 0.2$, $\rho = 0.5$, and $\rho = 0.7$. Note that, for the four methods, the *logistic* classifier is used to perform cross-validation.

Table 3
The classifiers involved in the illustrations.

abbreviation	classifier name	type	use a method of point prediction classifier?	parameters
ndc	non-deterministic	imprecise	yes	β
ncc	naive credal	imprecise	no	s
eclair rank 01	eclair relabelling: rank reasoning: 01 gain	imprecise	yes	none
eclair entropy 01	eclair relabelling: entropy reasoning: 01 gain	imprecise	yes	ρ
eclair rank PIG	eclair relabelling: rank reasoning: pignistic ndc	imprecise	yes	β
eclair entropy PIG	eclair relabelling: entropy reasoning: pignistic ndc	imprecise	yes	ρ β
eclair rank GFB	relabelling: rank reasoning: generalised F_β score	imprecise	yes	β
eclair entropy GFB	eclair relabelling: entropy reasoning: generalised F_β score	imprecise	yes	ρ β
nbc	naive Bayes	point prediction	-	-
knn	k-Nearest Neighbour	point prediction	-	-
eknn	evidential knn	point prediction	-	-
cart	decision tree	point prediction	-	-
rfc	random forest	point prediction	-	-
lda	linear discriminant analysis	point prediction	-	-
svm	support vector machine	point prediction	-	-
ann	artificial neural networks	point prediction	-	-
logistic	logistic	point prediction	-	-

Table 4
The new labels after relabelling.

	{a}	{b}	{a,b}	{c}	{a,c}	{b,c}	{a,b,c}
Rank method	481	482	31	494	0	12	0
Entropy method ($\rho = 0.2$)	430	392	132	465	0	69	12
Entropy method ($\rho = 0.5$)	468	444	61	482	0	28	17
Entropy method ($\rho = 0.7$)	481	480	0	493	0	2	44

The rank method relabels few samples compared to the entropy methods which can be explained by the fact that in most cases the obtained posterior probability gives the maximum value to the true class of the sample. In such cases the rank method does not relabel the samples whereas with the entropy methods, even if this situation is encountered, the samples are relabelled when the entropy is high. As one can expect, with the entropy methods a large number of samples are relabelled when ρ decreases. One can see in Fig. 1 that the relabelled samples with a subset of two classes constitute a large boundary between the two original classes. It is more obvious when $\rho = 0.2$ for the subsets $\{a, b\}$ and $\{b, c\}$. All the samples that are on the wrong side of the boundary are relabelled with the whole set $\{a, b, c\}$. In the case of a high entropy threshold, all the samples, where the posterior probability does not give the maximum value to the true class, are relabelled by the whole set, $\{a, b, c\}$ (see Equation (4)).

In order to evaluate the performance of the imprecise classifiers built from the considered training data, a dataset of 50 samples for each class label are generated using the same bivariate Gaussian distributions with a Gaussian noise $\mathcal{N}(\mu = (0, 0), \Sigma = 0.001 I_2)$. Note that, again the *logistic* classifier is used to provide the posterior probability distribution for *eclair rank*, *eclair entropy* and *ndc*. Moreover, several values of the parameters presented in Table 3 are tested and the results are shown in Table 5. Note that, for the considered testing data, the methods of point prediction obtain the following accuracies: *logistic*, *ann*: 94.67%; *svm*, *rfc*, *eknn*: 95.33%; *nbc*, *lda*, *cart*: 96%; and *knn*: 96.67%.

As one can see in Table 5, the imprecise classifiers have performances close to those of the point prediction classifiers for some fixed parameters but in these cases they are not cautious enough, the *eclair* and *ndc* performances are better

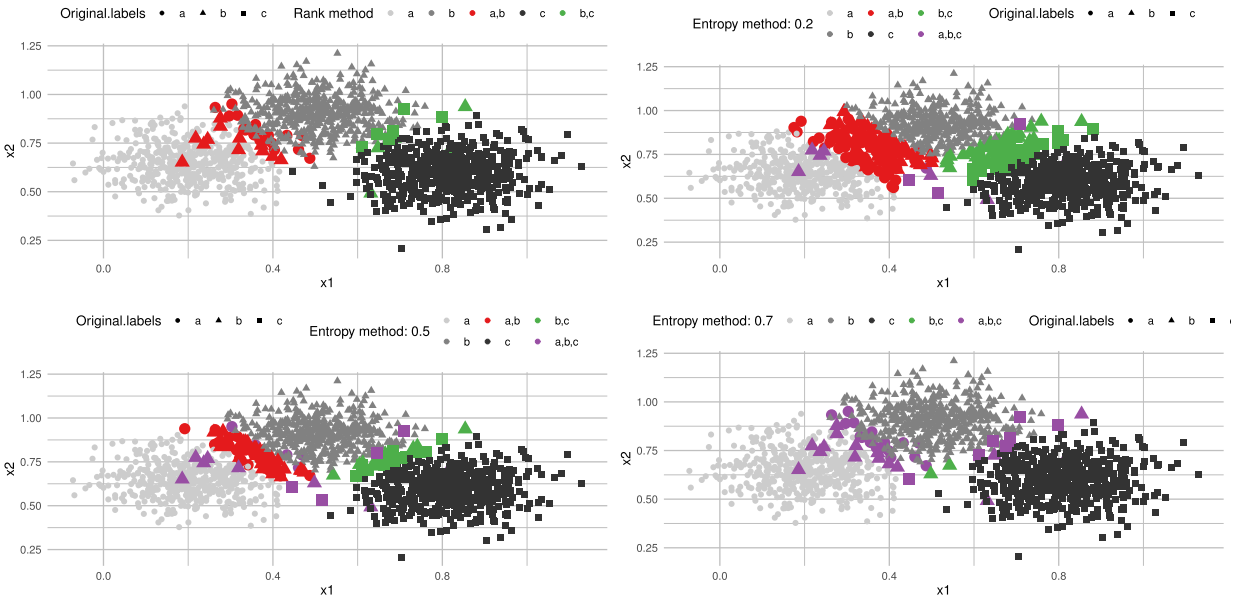

Fig. 1. Relabelling the training data using two methods with different parameters. (For interpretation of the colours in the figure(s), the reader is referred to the web version of this article.)

compared to those of *ncc*. Furthermore, *eclair* and *ndc* can reach very good performances for some values of β. For instance, *eclair rank GFB* and *eclair entropy GFB* have the best accuracy, u_{65} and u_{50} performances for small values of β while *ndc* has good imprecise accuracy and u_{80} performances. If one wants to be very cautious, it is better to choose *eclair entropy* classifiers as they can provide predictions with 100% good imprecise predictions for some parameters but in return they avoid precise predictions for several samples and then have a low accuracy performance. Generally, *ndc* and *eclair* classifiers give good average performances (see the last column of Table 5). This shows the trade-off that those imprecise classifiers are able to guarantee between cautiousness and efficiency.

To see the effect of these hyper-parameters more closely, Figs. 2 and 3 show the performances of the *eclair* classifiers related to two different point prediction methods and the hyper-parameters ρ and β using the same simulated data. Fig. 2 shows the results obtained when using the *svm* point prediction method while Fig. 3 shows the results obtained when using the *logistic* point prediction method. The first remark that can be made about the two figures is that regardless the performances of the two point prediction methods related to those data, the behaviours of the five performance measure curves are the same in the two figures. For that reason only Fig. 2 is commented. The first classifier studied is the *eclair entropy 01* classifier presented in the part "ENTR.01" in the Fig. 2 that has only the entropy threshold ρ as hyper-parameter ($\rho_1 = \rho$ and $\rho_2 = \rho - 0.05$). As expected, when ρ is very close to 0 the number of relabelled samples is very large and the size of the label subsets is also large thus the chance to contain the true class is close to 1, i.e. *imprecise accuracy* performance close to 100%, and the chance to predict the true class of a sample is low, i.e., accuracy performance is low. The performances measures u_{50}, u_{65}, u_{80} and the accuracy performance increase until ρ reaches a value ρ^* close to 0.25 while *imprecise accuracy* performance is rather constant. Above the value ρ^* all the performance measures curves become almost constant. Concerning the other three parts of Fig. 2 "RANK.PIG", "ENTR.PIG" and "ENTR.GFB" the hyper-parameter ρ is fixed ($\rho_1 = 0.25$ and $\rho_2 = 0.2$) and the hyper-parameter β is considered as variable. For those classifiers, the effect of the values of the hyper-parameter β on the performance measures is important. Indeed, with "ENTR.GFB" the imprecise accuracy reaches the performance close to 100% for values of β that are close to 0 while the accuracy decreases quickly. This behaviour is the same for the two other classifiers "RANG.PIG" and "ENTR.PIG" but the decrease and the increase in performance is lower. Thus the advantage of the "ENTR.GFB" classifier is to offer an optimal cautiousness with an accuracy very high, i.e., β is close to 0.

To highlight the good or bad predictions of the imprecise classifiers for some specific samples that are in overlapping or isolated regions, Fig. 4 shows the predictions for each sample of the test data provided by some imprecise classifiers. Ten samples among the 150 samples seem to be difficult to classify, i.e., the methods of point prediction fail to correctly classify them. The ten difficult samples are labelled by their numbers in Fig. 4. It can be noted that with $\beta = 0.1$ for the *eclair rank GFB* classifier, three samples have precise and correct predictions, five have precise and incorrect predictions, and two have "correct", i.e., contain the true class, imprecise predictions with two classes. The three other classifiers of Fig. 4 are more cautious but less efficient. The classifiers *ncc* ($s = 0.08$) and *ndc* ($\beta = 3$) provide less number of imprecise predictions than *eclair entropy* ($\rho = 0.5$ and $\beta = 1$). *eclair entropy* ($\rho = 0.5$ and $\beta = 1$) is the only one that does not provide incorrect predictions.

Table 5
The performances of the imprecise classifiers for different parameters.

		Performances					
	Parameters	accuracy	imprAcc	u_{80}	u_{65}	u_{50}	average
eclair rank 01	-	88.67	98.67	96.67	95.17	93.67	94.57
eclair rank PIG	$\beta \in [0, 0.17]$	**96**	96	96	96	**96**	96
	$\beta = 0.5$	94	97.33	96.67	96.17	95.67	95.97
	$\beta = 2.1$	86	98.67	96.13	94.23	92.33	93.47
	$\beta = 3$	83.33	98.67	95.6	93.3	91	92.38
eclair rank GFB	$\beta \in [0, 0.08]$	**96**	96	96	96	**96**	96
	$\beta = 0.1$	95.33	96.67	96.4	**96.2**	**96**	**96.12**
	$\beta = 0.5$	91.33	98	96.67	95.67	94.67	95.27
	$\beta = 2.1$	86	98.67	96.13	94.23	92.33	93.47
	$\beta = 3$	83.33	98.67	95.6	93.3	91	92.38
eclair entropy 01	$\rho = 0.2$	81.33	**100**	95.47	92.73	90	91.91
	$\rho = 0.5$	86.67	**100**	96.27	94.36	92.44	93.95
	$\rho = 0.6$	86.67	99.33	95.6	93.8	92	93.48
	$\rho = 0.7$	87.33	99.33	94.53	92.93	91.33	93.09
eclair entropy PIG	$\rho = 0.2$, $\beta \in [0, 0.09]$	95.33	95.33	95.33	95.33	95.33	95.33
	$\rho = 0.2$, $\beta = 0.5$	90	97.33	95.87	94.77	93.67	94.33
	$\rho = 0.5$, $\beta \in [0, 0.21]$	94.67	94.67	94.67	94.67	94.67	94.67
	$\rho = 0.5$, $\beta = 0.5$	91.33	97.33	96	95.11	94.22	94.8
	$\rho = 0.5$, $\beta = 1.3$	86	**100**	96.8	94.73	92.67	94.04
	$\rho = 0.5$, $\beta = 2$	81.33	**100**	95.07	92.37	89.67	91.69
eclair entropy GFB	$\rho = 0.2$, $\beta \in [0, 0.01]$	**96**	96	96	96	**96**	96
	$\rho = 0.2$, $\beta = 0.5$	88	99.33	96.67	95	93.33	94.47
	$\rho = 0.5$, $\beta \in [0, 0.1]$	94.67	94.67	94.67	94.67	94.67	94.67
	$\rho = 0.5$, $\beta = 1$	84.00	**100**	96	93.67	91.33	93
	$\rho = 0.5$, $\beta = 2$	76.67	**100**	93.2	89.88	86.56	89.26
	$\rho = 0.6$, $\beta = 0.2$	93.33	96.67	96	95.5	95	95.3
	$\rho = 0.7$, $\beta = 0.2$	94	95.33	94.8	94.62	94.44	94.64
ncc	$s = 0.1$	83.33	96.67	93.07	91.14	89.22	90.69
	$s = 0.08$	88.00	96.67	94.13	92.90	91.67	92.67
	$s = 0.05$	88.0	96.0	94.4	93.2	92.0	92.72
	$s = 0.01$	91.33	94.00	93.47	93.07	92.67	92.91
	$s = 0.005$	91.33	92.00	91.87	91.77	91.67	91.73
	$s = 0.0001$	92	92	92	92	92	92
ndc	$\beta \in [0, 0.36]$	94.67	94.67	94.67	94.67	94.67	94.67
	$\beta = 0.37$	94.67	95.33	95.20	95.1	95	95.06
	$\beta = 0.5$	94	96	95.6	95.3	95	95.18
	$\beta = 2.1$	92	98	96.67	95.78	94.89	95.47
	$\beta = 3$	90.67	99.33	**97.47**	96.18	94.89	95.71
	$\beta = 4$	88.67	99.33	97.07	95.48	93.89	94.89
	$\beta = 5$	88.67	99.33	96.93	95.36	93.78	94.81

4.2. Comparing the imprecise classifiers' performances using UCI data

The second illustration concerns the comparison of the performances of four imprecise classifiers: *ndc*, *ncc*, *eclair rank GFB* and *eclair entropy GFB* using 11 UCI data and the same performances measures used in the first illustration of Subsection 4.1 (see the definition in Section 2.4). A brief description of the selected UCI data is given in Table 6.

Each of those classifiers has hyper parameters that should be optimized and some of them use a method of point prediction. Table 7 presents a summary of those parameters, the steps in which they are involved and the sets from which they are selected. Note that the methods of point prediction are selected from the list presented in Table 3: *PRCL*={*nbc*, *knn*, *eknn*, *cart*, *rfc*, *lda*, *svm*, *ann* and *logistic* }.

As the aim is to build a classifier that offers the best trade-off between cautiousness and efficiency, the criterion that is used to optimize the parameter of the four classifier is the average of the five measures of performance: accuracy, imprecise accuracy, u_{80}, u_{65} and u_{50}. The experimentation procedure is conduct as follows. Each dataset is split randomly 10 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. Each cell in Table 8 presents the average performance of the classifiers on the test data of the 10 splits. As shown in Table 8, the most important observation is that even if some samples are classified with subsets of classes, the accuracy performance of the imprecise classifiers remains close to that of the methods of point prediction. The exception is the *ncc* classifier that has accuracy performances far from the best ones. Indeed, *ncc* is too cautious and in some situations, it has the best imprecise accuracy performances. It is the case for the *Wine* and *PID* data. As mentioned

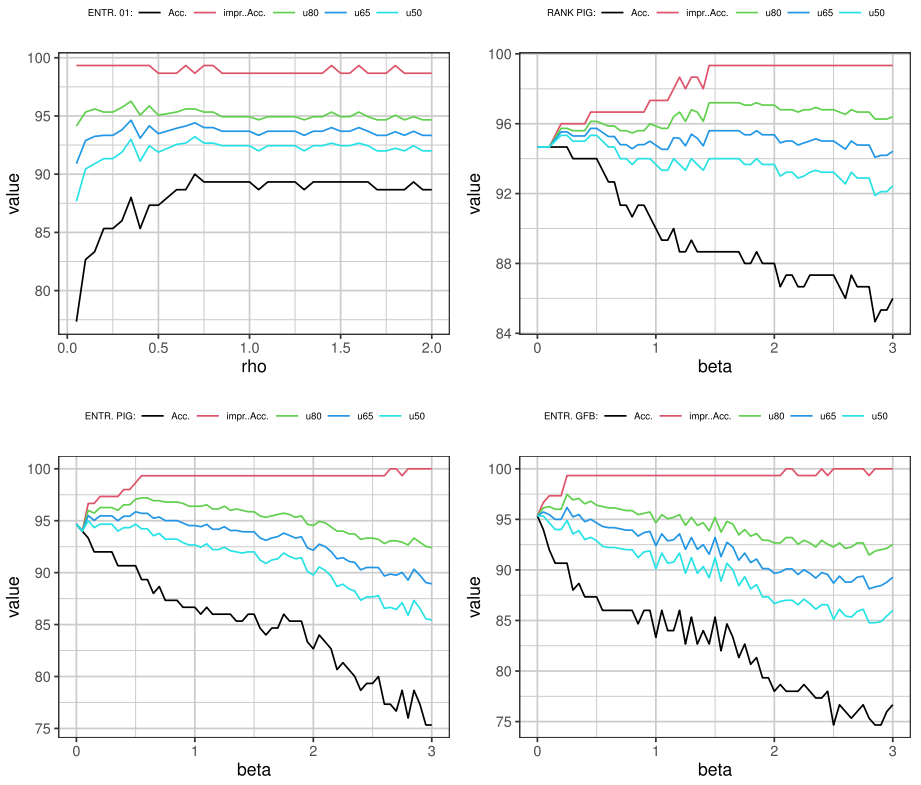

Fig. 2. *eclair* classifiers' performances using *svm* point prediction method. - ENTR.01: *eclair* classifier using the entropy relabelling method ($\rho_1 = \rho$, $\rho_2 = \rho - 0.05$) and the *01 gain* reasoning method. - RANK.PIG: *eclair* classifier using the rank relabelling method and the *pignistic ndc* reasoning method. - ENTR.PIG: *eclair* classifier using the entropy relabelling method ($\rho_1 = 0.25$, $\rho_2 = 0.2$) and the *pignistic ndc* reasoning method. - ENTR.GFB: *eclair* classifier using the entropy relabelling method ($\rho_1 = 0.25$, $\rho_2 = 0.2$) and the *generalised* F_β *score* reasoning method. (For interpretation of the colours in the figure(s), the reader is referred to the web version of this article.)

Table 6
Information about the considered UCI data.

nom	# instances	# inputs	# class	abbreviation
Iris	150	4	3	Iris
Breast Cancer	683	9	2	BC
Wine	178	13	3	Wine
Ionosphere	351	32	2	IS
Diabetes	145	5	3	DBT
Glass	214	9	6	Glass
Pima Indians Diabetes	392	9	2	PID
Sonar	208	60	2	Sonar
Seeds	210	7	3	Seeds
Forest	523	27	4	Forest
Ecoli	327	5	5	Ecoli

Table 7
The hyper parameters involved in the models of the imprecise classifiers.

classifier	relabelling	posterior probabilities	reasoning step
ndc	-	$cl \in PRCL$	$\beta \in [0, 3]$
ncc	-	IDM ($s \in [10^{-10}, 2]$)	-
eclair rank GFB	$cl \in PRCL$	*cl*	$\beta \in [0, 3]$
eclair entropy GFB	$cl \in PRCL$ $\rho \in [0.1, 1]$	*cl*	$\beta \in [0, 3]$

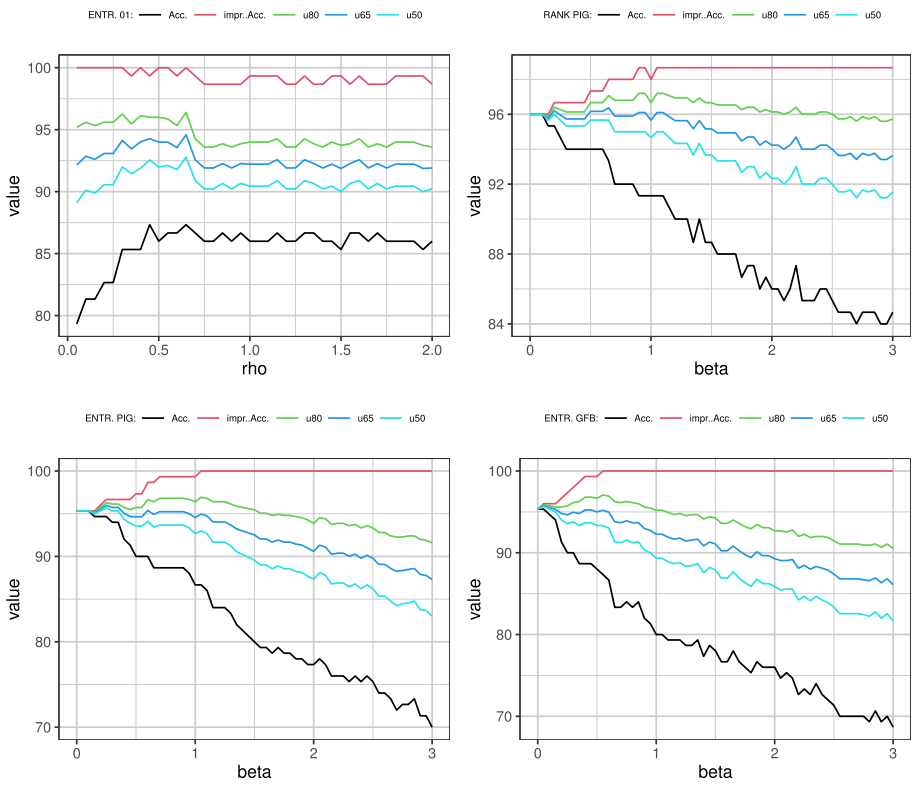

Fig. 3. *eclair* classifiers' performances using *logistic* point prediction method. - ENTR.01: *eclair* classifier using the entropy relabelling method ($\rho_1 = \rho$, $\rho_2 = \rho - 0.05$) and the *01 gain* reasoning method. - RANK.PIG: *eclair* classifier using the rank relabelling method and the *pignistic ndc* reasoning method. - ENTR.PIG: *eclair* classifier using the entropy relabelling method ($\rho_1 = 0.25$, $\rho_2 = 0.2$) and the *pignistic ndc* reasoning method. - ENTR.GFB: *eclair* classifier using the entropy relabelling method ($\rho_1 = 0.25$, $\rho_2 = 0.2$) and the *generalised* F_β *score* reasoning method. (For interpretation of the colours in the figure(s), the reader is referred to the web version of this article.)

in the case of the simulated data of Subsection 4.1, *eclair* classifiers obtain accuracy scores close to the best ones of the methods of point prediction and the imprecise predictions of the difficult samples are all almost as good for several datasets. This can be seen for the following data: *Iris*, *BC*, *IS*, *DBT*, *Seeds*, *Forest* and *Ecoli*. Generally, the *ndc* and *eclair* classifiers obtain close results except for the u_{50} measure where *ndc* obtains slightly better performances.

Furthermore, we can also compare these results to those of the imprecise classifier named *preorder* in [19]. As it is mentioned in the introduction Section, this classifier constructs a binary relation on the set of classes and the subset of the non-dominated classes is considered as the prediction for the new sample. The dominance relation is based on different quantifications of uncertainties. The authors conducted the experiments for some UCI data under the same conditions as those considered in our experiments but only two measures are used: u_{80} and u_{65}. Table 9 presents the comparisons between the performances of *preorder* reported in [19] and the performances of *eclair* classifiers and shows that except the "Forest" data, the *eclair* classifiers have the best performances on the other five datasets. In addition, we tested the five measures with a version of *eclair* where the training step is performed using the evidential *k*-nearest neighbourhoods EKNN with partially labelled samples [46][47]. Unfortunately, the obtained performances of this version fall far short of what is measured for the other versions of the *eclair* classifiers.

5. Conclusion

This paper proposes an approach for imprecise classification based on an imprecise relabelling of the training data and a generalisation of the F_β score within the framework of belief functions. Several imprecise classifiers can be built from the approach depending on the choice of the method of relabelling and the choice of the gain function involved in the reasoning step. Some choices appear preferable as they have less hyper parameters, i.e., *eclair rank 01*, but they do not obtain the best performances. Two illustrations are proposed to compare the proposed approach to the state-of-the-art approaches. First, simulated data are used to show how those classifiers deal with difficult samples. The *eclair* classifiers offer the best compromise between cautiousness and efficiency on this data. Second, the comparisons are conducted using the UCI data. The results show that the *eclair* classifiers have accuracy performances close to those of the point prediction classifiers while their performances on the other measures of cautiousness remain very competitive. Concerning the hyper parameters, the *eclair* classifiers can be improved by making them free from the methods of point prediction. However, the

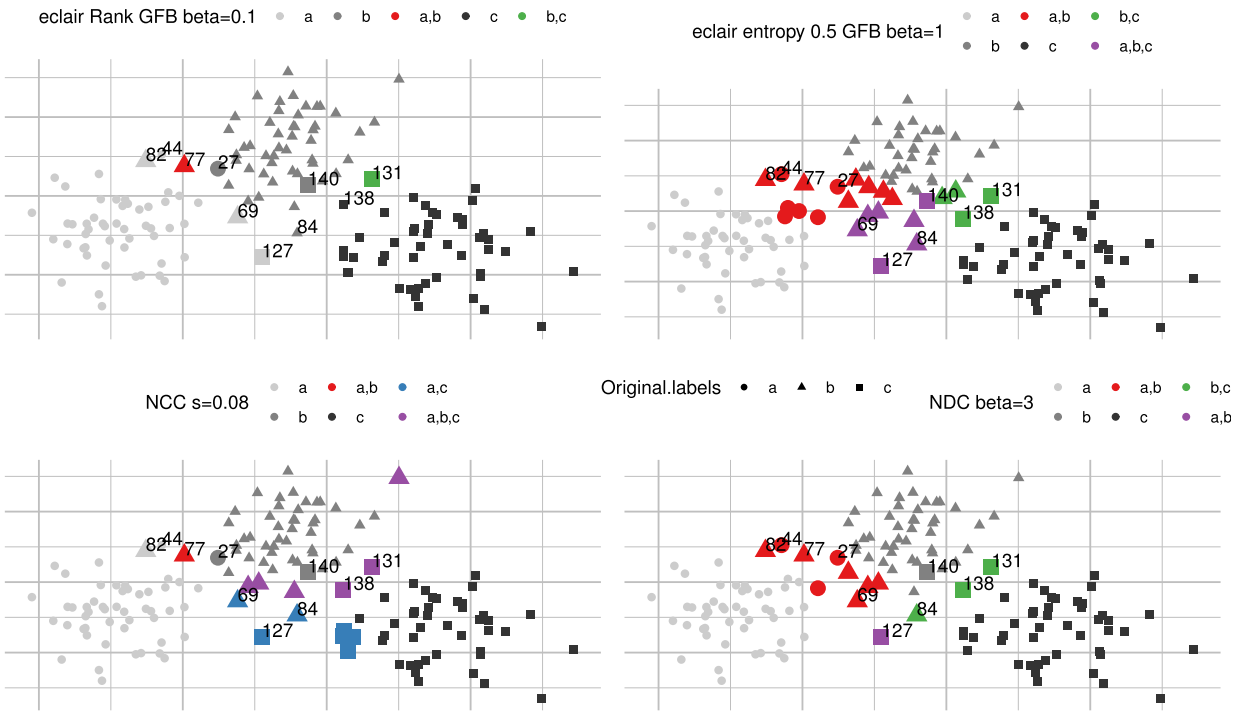

Fig. 4. The predictions of four selected imprecise classifiers: a large size is given to the points symbols representing predictions that are errors or imprecise. (For interpretation of the colours in the figure(s), the reader is referred to the web version of this article.)

Table 8
The imprecise classifiers' performances on the UCI data.

		Iris	BC	Wine	IS	DBT	Glass	PID	Sonar	Seeds	Forest	Ecoli
Accuracy	svm	95.67	96.15	98.53	**94.57**	85	68.5	75.51	**84.15**	94	90.19	90
	rfc	96	**97.19**	98.24	93.86	**98.93**	**77**	79.10	83.9	92.62	91.17	**90.16**
	lda	**97.67**	95.85	**99.41**	85.29	91.07	62.5	79.1	70.73	**97.38**	89	90.62
	nbc	96	95.93	97.35	83.71	93.21	38	77.69	67.32	92.86	87.28	88.91
	knn	96.33	96.59	65.59	83.71	91.07	61.5	74.36	63.66	90.48	90	89.53
	cart	94	94.81	91.76	89	97.86	70.25	77.18	73.17	92.38	87.96	84.69
	ann	96.33	94.67	90	86.14	95.36	64.75	78.08	80	95.48	86.89	81.09
	eknn	97	96.67	72.06	92.86	92.14	67	72.95	81.46	87.62	**91.55**	89.84
	logistic	96	96.59	93.82	83.43	96.07	64.25	**79.62**	71.71	95.71	87.77	89.53
	eclair rank GFB	96	94.6	97.35	92.28	98.21	65.25	70.51	67.32	94.04	89.7	88.12
	eclair entropy GFB	96	96.22	97.94	91.71	95.71	52.5	66.8	73	94.04	88.44	87.34
	ndc	96.67	96.51	97.64	94.28	**98.93**	72	73.84	76.58	96.66	89.41	88.12
	ncc	91	95.55	88.82	62.57	87.85	23.25	15.76	30.97	83.57	24.85	38.28
imprAcc	eclair rank GFB	96	**97.85**	97.35	94.57	**98.93**	78.25	84.49	**94.88**	97.86	90.98	90.62
	eclair entropy GFB	**98.33**	96.88	97.94	**95.57**	98.21	77.5	86.02	89.51	**98.1**	90.87	89.53
	ndc	96.67	96.88	97.94	95.42	**98.93**	**80.5**	82.43	88.3	97.85	**91.84**	**92.5**
	ncc	96.67	96	**99.11**	87.14	95.35	65.5	**95.38**	90.48	94.76	87.57	85
*u*80	eclair rank GFB	96	**97.2**	97.35	94.11	98.64	75.16	81.7	**89.36**	97.1	90.65	89.8
	eclair entropy GFB	**97.86**	96.75	**97.94**	94.8	97.21	69.1	**82.18**	86.2	96.57	90.32	89.09
	ndc	96.67	96.81	97.88	**95.2**	**98.93**	**78.75**	80.71	85.95	**97.62**	**91.26**	**91.62**
	ncc	95.4	95.91	95.47	82.2	93.57	39.84	79.46	78.58	92.33	71.23	72.24
*u*65	eclair rank GFB	96	**96.75**	97.35	93.78	98.54	73.27	**79.6**	**85.23**	96.52	90.47	89.47
	eclair entropy GFB	**97.51**	96.65	**97.94**	94.22	96.88	65.82	79.3	83.7	96.02	89.96	88.76
	ndc	96.67	**96.75**	97.83	**95.02**	**98.93**	**77.48**	79.43	84.19	**97.44**	**90.91**	**90.96**
	ncc	94.56	95.84	94.05	78.54	92.47	35.88	67.51	69.65	90.67	62.20	65.58
*u*50	eclair rank GFB	96	96.2	97.35	93.43	98.45	71.37	77.5	81.1	95.95	90.3	89.14
	eclair entropy GFB	**97.16**	96.55	**97.94**	93.64	96.54	62.54	76.41	81.22	95.47	89.61	88.43
	ndc	96.67	**96.7**	97.79	**94.85**	**98.93**	**76.2**	**78.14**	**82.44**	**97.26**	**90.55**	**90.31**
	ncc	93.72	95.77	92.64	74.85	91.36	31.93	55.57	60.73	89	53.17	58.92

Table 9
Comparison of the performances of *preorder* classifier reported in [19] and *eclair* classifiers' performances on some UCI data.

		Iris	Wine	Glass	Seeds	Forest	Ecoli
$u80$	eclair rank GFB	96	97.35	**75.16**	**97.1**	90.65	**89.8**
	eclair entropy GFB	**97.86**	**97.94**	69.1	96.57	90.32	89.09
	preorder	90.45	95.89	67.32	92.15	**92.15**	80.66
$u65$	eclair rank GFB	96	97.35	**73.27**	**96.52**	**90.47**	**89.47**
	eclair entropy GFB	**97.51**	**97.94**	65.82	96.02	89.96	88.76
	preorder	83.29	93.18	57.24	88.16	88.82	75.25

other hyper parameters are required because during the relabelling step they enable the detection of difficult samples in the training data and those involved in the reasoning step enable control of the trade-off between cautiousness and efficiency. Those parameters can be considered as user-control parameters whose values depend on the targeted application.

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.

References

[1] E. Hüllermeier, Learning from imprecise and fuzzy observations: data disambiguation through generalized loss minimization, Int. J. Approx. Reason. 55 (2014) 1519–1534.
[2] I. Couso, D. Dubois, Statistical reasoning with set-valued information: ontic vs. epistemic views, Int. J. Approx. Reason. 55 (2014) 1502–1518.
[3] L. Jacquin, A. Imoussaten, F. Trousset, D. Perrin, J. Montmain, Control of waste fragment sorting process based on mir imaging coupled with cautious classification, Resour. Conserv. Recycl. (2020) 105258.
[4] H. Xiong, M. Li, T. Jiang, S. Zhao, Classification algorithm based on nb for class overlapping problem, Appl. Math. 7 (2013) 409–415.
[5] H.K. Lee, S.B. Kim, An overlap-sensitive margin classifier for imbalanced and overlapping data, Expert Syst. Appl. 98 (2018) 72–83.
[6] R.A. Fisher, The fiducial argument in statistical inference, Ann. Eugenics 6 (1935) 391–398.
[7] J. Neyman, X-outline of a theory of statistical estimation based on the classical theory of probability, Philos. Trans. R. Soc. Lond. Ser. A, Math. Phys. Sci. 236 (1937) 333–380.
[8] C.-K. Chow, An optimum character recognition system using decision functions, IRE Trans. Electron. Comput. (1957) 247–254.
[9] C. Chow, On optimum recognition error and reject tradeoff, IEEE Trans. Inf. Theory 16 (1970) 41–46.
[10] T.M. Ha, The optimum class-selective rejection rule, IEEE Trans. Pattern Anal. Mach. Intell. 19 (1997) 608–615.
[11] J.J.d. Coz, J. Díez, A. Bahamonde, Learning nondeterministic classifiers, J. Mach. Learn. Res. 10 (2009) 2273–2293.
[12] V. Vovk, A. Gammerman, G. Shafer, Conformal prediction, in: Algorithmic Learning in a Random World, 2005, pp. 17–51.
[13] H. Papadopoulos, Inductive conformal prediction: theory and application to neural networks, in: Tools in Artificial Intelligence, Citeseer, 2008.
[14] L. Ma, T. Denoeux, Partial classification in the belief function framework, Knowl.-Based Syst. (2021) 106742.
[15] M. Zaffalon, Statistical inference of the naive credal classifier, in: ISIPTA, Vol. 1, 2001, pp. 384–393.
[16] M. Zaffalon, A credal approach to naive classification, in: ISIPTA, Vol. 99, 1999, pp. 405–414.
[17] J.-M. Bernard, An introduction to the imprecise Dirichlet model for multinomial data, Int. J. Approx. Reason. 39 (2005) 123–150.
[18] L.M. De Campos, J.F. Huete, S. Moral, Probability intervals: a tool for uncertain reasoning, Int. J. Uncertain. Fuzziness Knowl.-Based Syst. 2 (1994) 167–196.
[19] V.-L. Nguyen, S. Destercke, M.-H. Masson, E. Hüllermeier, Reliable multi-class classification based on pairwise epistemic and aleatoric uncertainty, in: International Joint Conference on Artificial Intelligence (IJCAI-18), 2018, pp. 5089–5095.
[20] R. Senge, S. Bösner, K. Dembczyński, J. Haasenritter, O. Hirsch, N. Donner-Banzhoff, E. Hüllermeier, Reliable classification: learning classifiers that distinguish aleatoric and epistemic uncertainty, Inf. Sci. 255 (2014) 16–29.
[21] B. Quost, M.-H. Masson, S. Destercke, Dealing with atypical instances in evidential decision-making, in: International Conference on Scalable Uncertainty Management, Springer, 2020, pp. 217–225.
[22] J. Fürnkranz, E. Hüllermeier, Preference learning and ranking by pairwise comparison, in: Preference Learning, Springer, 2010, pp. 65–82.
[23] E. Hüllermeier, K. Brinker, Learning valued preference structures for solving classification problems, Fuzzy Sets Syst. 159 (2008) 2337–2352.
[24] G. Shafer, A Mathematical Theory of Evidence, Vol. 42, Princeton University Press, 1976.
[25] C.N. Silla, A.A. Freitas, A survey of hierarchical classification across different application domains, Data Min. Knowl. Discov. 22 (2011) 31–72.
[26] K. Punera, J. Ghosh, Enhanced hierarchical classification via isotonic smoothing, in: Proceedings of the 17th International Conference on World Wide Web, ACM, 2008, pp. 151–160.
[27] A. Binder, M. Kawanabe, U. Brefeld, Efficient classification of images with taxonomies, in: Asian Conference on Computer Vision, Springer, 2009, pp. 351–362.
[28] S. Harispe, A. Imoussaten, F. Trousset, J. Montmain, On the consideration of a bring-to-mind model for computing the information content of concepts defined into ontologies, in: Fuzzy Systems (FUZZ-IEEE), 2015 IEEE International Conference on, IEEE, 2015, pp. 1–8.
[29] M. Ceci, Hierarchical text categorization in a transductive setting, in: 2008 IEEE International Conference on Data Mining Workshops, IEEE, 2008, pp. 184–191.
[30] T. Denoeux, Maximum likelihood estimation from uncertain data in the belief function framework, IEEE Trans. Knowl. Data Eng. 25 (2013) 119–130.
[31] O. Kanjanatarakul, S. Sriboonchitta, T. Denoeux, Forecasting using belief functions: an application to marketing econometrics, Int. J. Approx. Reason. 55 (2014) 1113–1128.
[32] O. Kanjanatarakul, T. Denœux, S. Sriboonchitta, Prediction of future observations using belief functions: a likelihood-based approach, Int. J. Approx. Reason. 72 (2016) 71–94.
[33] T. Denoeux, A k-nearest neighbor classification rule based on dempster-shafer theory, IEEE Trans. Syst. Man Cybern. 25 (1995) 804–813.
[34] N. Sutton-Charani, A. Imoussaten, S. Harispe, J. Montmain, Evidential bagging: combining heterogeneous classifiers in the belief functions framework, in: International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems, Springer, 2018, pp. 297–309.

[35] D. Alshamaa, F.M. Chehade, P. Honeine, A hierarchical classification method using belief functions, Signal Process. 148 (2018) 68–77.
[36] M.P. Naeini, B. Moshiri, B.N. Araabi, M. Sadeghi, Learning by abstraction: hierarchical classification model using evidential theoretic approach and bayesian ensemble model, Neurocomputing 130 (2014) 73–82.
[37] J. Abellan, A.R. Masegosa, Imprecise classification with credal decision trees, Int. J. Uncertain. Fuzziness Knowl.-Based Syst. 20 (2012) 763–787.
[38] G. Yang, S. Destercke, M.-H. Masson, The costs of indeterminacy: how to determine them?, IEEE Trans. Cybern. 47 (2016) 4316–4327.
[39] G. Tsoumakas, I. Vlahavas, Random k-labelsets: an ensemble method for multilabel classification, in: European Conference on Machine Learning, Springer, 2007, pp. 406–417.
[40] M. Zaffalon, G. Corani, D. Mauá, Evaluating credal classifiers by utility-discounted predictive accuracy, Int. J. Approx. Reason. 53 (2012) 1282–1301.
[41] S. Kanj, F. Abdallah, T. Denoeux, K. Tout, Editing training data for multi-label classification with the k-nearest neighbor rule, Pattern Anal. Appl. 19 (2016) 145–161.
[42] S. Lallich, F. Muhlenbach, D.A. Zighed, Improving classification by removing or relabeling mislabeled instances, in: International Symposium on Methodologies for Intelligent Systems, Springer, 2002, pp. 5–15.
[43] F. Muhlenbach, S. Lallich, D.A. Zighed, Identifying and handling mislabelled instances, J. Intell. Inf. Syst. 22 (2004) 89–109, https://doi.org/10.1023/A:1025832930864.
[44] L. Jiao, T. Denœux, Q. Pan, Evidential editing k-nearest neighbor classifier, in: European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty, Springer, 2015, pp. 461–471.
[45] J. Zhang, S. Subasingha, K. Premaratne, M.-L. Shyu, M. Kubat, K. Hewawasam, A novel belief theoretic association rule mining based classifier for handling class label ambiguities, in: Proc. Workshop Foundations of Data Mining (FDM'04), Int. Conf. Data Mining (ICDM'04), 2004.
[46] L.M. Zouhal, T. Denoeux, An evidence-theoretic k-nn rule with parameter optimization, IEEE Trans. Syst. Man Cybern., Part C, Appl. Rev. 28 (1998) 263–271.
[47] E. Côme, L. Oukhellou, T. Denoeux, P. Aknin, Learning from partially supervised data using mixture models and belief functions, Pattern Recognit. 42 (2009) 334–348.
[48] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, Smote: synthetic minority over-sampling technique, J. Artif. Intell. Res. 16 (2002) 321–357.