A K-nearest neighbours method based on lower previsions

Sebastien Destercke

INRA/CIRAD, UMR1208, 2 place P. Viala, F-34060 Montpellier cedex 1, France
sebastien.destercke@supagro.inra.fr

Abstract. K-nearest neighbours algorithms are among the most popular existing classification methods, due to their simplicity and good performances. Over the years, several extensions of the initial method have been proposed. In this paper, we propose a K-nearest neighbours approach that uses the theory of imprecise probabilities, and more specifically lower previsions. This approach handles very generic models when representing imperfect information on the labels of training data, and decision rules developed within this theory allows to deal with issues related to the presence of conflicting information or to the absence of close neighbours. We also show that results of the classical voting K-NN procedures and distance-weighted k-NN procedures can be retrieved.

Keywords: Classification, lower prevision, nearest neighbours.

1 Introduction

The k-nearest neighbours (K-NN) classification procedure is an old rule [1] that uses the notion of similarity and distance with known instances to classify a new one. Given a vector $\boldsymbol{x} \in \mathbb{R}^D$ of input features, a distance $d : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}$ and a data set of training samples composed of N couples $(\boldsymbol{x}_i, y_i)$ where $\boldsymbol{x}_i \in \mathbb{R}^D$ are feature values and $y_i \in \mathcal{Y} = \{\omega_1, \ldots, \omega_M\}$ is the class to which belongs the i^{th} sample, the voting k-NN procedure consists in choosing as the class y of $\boldsymbol{x}$ the one that is in majority in the k nearest neighbours.

One of the main drawback of the original algorithm is that it assumes that the k-nearest neighbors are relatively close to the instance to classify, and can act as reliable instances to estimate some conditional densities. It also assumes that all classes or patterns are well represented in the input feature space, and that the space is well sampled. In practice, this is rarely true, and the distance between a new instance and its nearest neighbour can be large. This makes the way basic k-NN procedure treats the training samples questionable Also, some classes of training samples may only be imperfectly known, and this uncertainty should be taken into account.

To integrate these various features, many extensions of the initial method have been proposed: use of weights to account for distance between neighbours and instance to classification [2]; use of distance and ambiguity rejection, to cope respectively with nearest neighbours whose distance from the instance to classify is too large and with nearest neighbours giving conflicting information [3]; use of uncertainty representations such as belief functions to cope with uncertainty [4]. For a detailed survey of the k-NN algorithm and its different extensions, see [5, Chap. 2].

As far as uncertainty representations are concerned, it can be argued that belief functions do not allow to model precisely all kinds of uncertainties. For example, they are unable to model exactly uncertainty given by probability intervals (i.e., lower and upper probabilistic bounds given on each class). Imprecise probability theory and walley's lower previsions [6] are uncertainty models that encompass belief functions as special cases. In this sense, they are more general and allow for a finer modelling of uncertainty.

In this paper, we propose and discuss a k-NN rule based on the use of Walley's lower prevision [6,7], and of the theory underlying them. As for the TBM k-NN procedure (based on evidence theory and on Dempster's rule of combintion), it allows to treat all the issues mentioned above without introducing any other parameters than the weights on nearest neighbours, however it does so with a different approach (being based on different calculus) and allows the use of more general uncertainty models than the TBM. In particular, we argue that using decision rules proper to the lower previsions approach allows to take account of ambiguities and distances without having to include additional parameters. Using these imprecise decision rules, we also introduce a criteria allowing to pick the "best" number k of nearest neighbours, balancing imprecision and accuracy. After recalling the material concerning lower previsions (Section 2) needed in this paper, we details the proposed method and its properties (Section 3), before finishing with some experiments (Section 4).

2 Lower previsions

This section introduces the very basics about lower previsions and associated tools needed in this paper. We refer to Miranda [7] and Walley [6] for more details.

2.1 Basics of lower previsions

In this paper, we consider that information regarding a variable X assuming its values on a (finite) space $\mathcal{X}$ counting N exclusive and disjoint elements is modelled by the means of a so-called coherent lower previsions. We denote by $\mathcal{L}(\mathcal{X})$ the set of real-valued bounded functions on $\mathcal{X}$. A lower prevision $\underline{P} : \mathcal{K} \to \mathbb{R}$ is a real-valued mapping on a subset $\mathcal{K} \subseteq \mathcal{L}(\mathcal{X})$. Given a lower prevision, the dual notion of upper prevision $\overline{P}$ is defined on the set $-\mathcal{K} = \{-f | f \in \mathcal{K}\}$ and is such that $\underline{P}(f) = -\overline{P}(-f)$. As discussed by Walley [6], lower previsions can be used to model information about the variable X. He interprets $\underline{P}(f)$ as the supremum buying price for the uncertain reward f.

Given a set $A \subseteq \mathcal{X}$, its lower probability $\underline{P}(A)$ is the lower prevision of its indicator function $\mathbf{1}_{(A)}$, that takes value one on A and zero elsewhere. The upper probability $\overline{P}(A)$ of A is the upper prevision of $\mathbf{1}_{(A)}$, and by duality $\underline{P}(A) = 1 - \overline{P}(A^c)$. To a lower prevision $\underline{P}$ can be associated a convex set $\mathcal{P}_{\underline{P}}$ of probabilities, such that

$$\mathcal{P}_{\underline{P}} = \{p \in \mathbb{P}_{\mathcal{X}} | (\forall f \in \mathcal{K})(E_p(f) \geq \underline{P}(f))\}$$

with $\mathbb{P}_{\mathcal{X}}$ the set of all probability mass functions over $\mathbb{P}_{\mathcal{X}}$ and $E_p(f) = \sum_{x \in \mathcal{X}} p(x)f(x)$ the expected value of f given p. As often done, $\mathcal{P}_{\underline{P}}$ will be called the credal set of $\underline{P}$.

A lower prevision is said to avoid sure loss iff $\mathcal{P}_{\underline{P}} \neq \emptyset$ and to be coherent iff it avoids sure loss and $\forall f \in \mathcal{K}$, $\underline{P}(f) = \min\{E_p(f)|p \in \mathcal{P}_{\underline{P}}\}$, i.e. iff $\underline{P}$ is the lower envelope of $\mathcal{P}_{\underline{P}}$. If a lower (upper) prevision is coherent, it corresponds to the lower (upper) expectation of $\mathcal{P}_{\underline{P}}$. If a lower prevision $\underline{P}$ avoids sure loss, its natural extension $\underline{E}(g)$ to a function $g \in \mathcal{L}(\mathcal{X})$ is defined as $\underline{E}(g) = \min\{E_p(g)|p \in \mathcal{P}_{\underline{P}}\}$. Note that $\underline{P}$ and its natural extension $\underline{E}$ coincide on $\mathcal{K}$ only when $\underline{P}$ is coherent, otherwise $\underline{P} \leq \underline{E}$ and $\underline{P}(f) < \underline{E}(f)$ for at least one f.

Lower previsions are very general uncertainty models, in that they encompass (at least from a static viewpoint) most of the other known uncertainty models. In particular both necessity measures of possibility theory [8] and belief measures of evidence theory [9] can be seen as particular lower previsions.

2.2 Vacuous mixture and lower previsions merging

When multiple sources provide possibly unreliable lower previsions modelling their beliefs, we must provide rules both to take this unreliability into account and to merge the different lower previsions into a single one, representing our final beliefs.

An extreme case of coherent lower prevision is the vacuous prevision $\underline{P}_v$ and its natural extension $\underline{E}_v$, which are such that $\underline{E}_v(g) = \inf_{\omega\in\mathcal{X}} g(\omega)$. It represents a state of total ignorance about the real value of X. Given a coherent lower prevision $\underline{P}$, its natural extension $\underline{E}$ and a scalar $\epsilon \in [0, 1]$, the (coherent) lower prevision $\underline{P}_\epsilon$ that we call vacuous mixture is such that $\underline{P}_\epsilon = \epsilon\underline{P} + (1 - \epsilon)\underline{P}_v$. Its natural extension $\underline{E}_\epsilon$ is such that $\underline{E}_\epsilon(f) = \epsilon\underline{E}(f) + (1 - \epsilon)\inf_{\omega\in\mathcal{X}} f(\omega)$, for any $f \in \mathcal{L}(\mathcal{X})$ and with $\underline{E}$ the natural extension of $\underline{P}$. ϵ can be interpreted as the probability that the information $\underline{P}$ is reliable, $1 - \epsilon$ being the probability of being ignorant. The vacuous mixture is a generalise both the the well-known linear-vacuous mixture and the classical discounting rule of belief functions. In terms of credal sets, it is equivalent to compute $\mathcal{P}_{\underline{P}_\epsilon}$ such that $\mathcal{P}_{\underline{P}_\epsilon} = \{\epsilon p_{\underline{P}} + (1 - \epsilon)p_v | p_{\underline{P}} \in \mathcal{P}_{\underline{P}}, p_v \in \mathbb{P}_{\mathcal{X}}\}$.

Now, if we consider k coherent lower previsions $\underline{P}_1, \ldots, \underline{P}_k$ and their natural extensions $\underline{E}_1, \ldots, \underline{E}_k$, then we can average them into a natural extension $\underline{E}_\sigma$ by merging them through an arithmetic mean, that is by considering $\underline{E}_\sigma(f) = \frac{1}{k}\sum_{i=1}^{k} \underline{E}_i(f)$ for any $f \in \mathcal{L}(\mathcal{X})$. This rule has been justified and used by different authors to merge coherent lower previsions or, equivalently, convex sets of probabilities [10].

2.3 Decision rules

Given some beliefs about a (finite) variable X and a set of preferences, the goal of decision rules is here to select the optimal values X can assume, i.e. the class to which X may belong. Here, we assume that preferences are modeled, for each $\omega \in \mathcal{X}$, by cost functions f'_ω, that is $f'_\omega(\omega')$ is the cost of selecting ω' when ω is the true class. When uncertainty over $\mathcal{X}$ is represented by a single probability p, the optimal class is the one whose expected cost is the lowest, i.e. $\widehat{\omega} = \arg\min_{\omega\in\mathcal{X}} E_p(f'_\omega)$, thus taking minimal risks. If the beliefs about the value of X are given by a lower prevision $\underline{P}$, the classical expected cost based decision has to be extended [11].

One way to do so is to still require the decision to be a single class. The most well-known decision rule in this category is the maximin rule, for which the final decision is such that

$$\widehat{\omega} = \arg\min_{\omega \in \mathcal{X}} \overline{E}_p(f'_\omega)$$

this amounts to minimising the upper expected cost, i.e., the worst possible consequence, and corresponds to a cautious decision. Other possible rules include minimising the lower expected cost or minimising a value in-between.

The other way to extend expected cost is to give as decision a set (possibly, but not necessarily reduced to a singleton) of classes, reflecting our indecision and the imprecision of our beliefs. This requires to build, among the possible choices (here, the classes), a partial ordering, and then to select only the choices that are not dominated by another one. Two such extensions are the interval ordering $\leq_I$ and the maximality ordering $\leq_M$. Using interval ordering, a choice ω is dominated by a choice ω', denoted by $\omega \leq_I \omega'$, iff $\overline{E}(f'_\omega) \leq \underline{E}(f_\omega)$, that is if the upper expected cost of picking ω' is sure to be lower than the lower expected cost of picking ω. The decision set $\widehat{\Omega}_I$ is then

$$\widehat{\Omega}_I = \{\omega \in \mathcal{X} | \; \nexists \omega' \text{s.t.} \omega \leq_I \omega'\}.$$

Using maximality ordering, a choice ω is dominated by a choice ω', denoted by $\omega \leq_M \omega'$, iff $\underline{E}(f_\omega - f_{\omega'}) > 0$. This has the following interpretation: given our beliefs, exchanging ω for ω' would have a strictly positive expected cost, hence we are not ready to do so. The decision set $\widehat{\Omega}_M$ is then

$$\widehat{\Omega}_M = \{\omega \in \mathcal{X} | \; \nexists \omega' \text{s.t.} \omega \leq_M \omega'\}.$$

The maximility ordering refines the Interval ordering and is stronger, in the sense that we always have $\widehat{\Omega}_M \subseteq \widehat{\Omega}_I$. Using these decision rules, the more precise and non-conflicting our information is, the smaller is the set of possible classes $\widehat{\Omega}$.

3 The method

Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ be N D-dimensional training samples, $\mathcal{Y} = \{\omega_1, \ldots, \omega_M\}$ the set of possible classes, and $\underline{P}_i : \mathcal{L}(\mathcal{Y}) \to [0, 1]$ be the lower prevision modelling our knowledge about the class to which the sample $\boldsymbol{x}_i$ belongs. Given a new instance $\boldsymbol{x}$ to classify, that is to which we have to assign a class $y \in \mathcal{Y}$, we denote by $\boldsymbol{x}_{(1)}, \ldots, \boldsymbol{x}_{(k)}$ its k ordered nearest neighbours (i.e. $d_{(i)} < d_{(j)}$ if $i \leq j$). For a given nearest neighbour $\boldsymbol{x}_{(i)}$, the knowledge $\underline{P}_{(i)}$ can be regarded as a piece of evidence related to the unknown class of $\boldsymbol{x}$. However, this piece of knowledge is not 100% reliable, and should be discounted by a value $\epsilon_i \in [0, 1]$ depending of its class, such that, for any $f \in \mathcal{L}(\mathcal{Y})$,

$$\underline{E}_{(i),\boldsymbol{x}}(f) = \epsilon_{(i)} \underline{E}_{(i)} + (1 - \epsilon_{(i)}) \inf_{\omega \in \mathcal{Y}} f(\omega).$$

It seems natural to ask for ϵ be a decreasing function of $d_{(i)}$, since the further away is the neighbour, the less reliable is the information it provides about the unknown class. Similarly to Denoeux proposal, we can consider the general formula

$$\epsilon = \epsilon_0 \phi(d_{(i)}),$$

where ϕ is a non-increasing function that can be depended of the class given by $\boldsymbol{x}_{(i)}$. In addition, the following conditions should hold:

$$0 < \epsilon_0 < 1 \quad ; \quad \phi(0) = 1 \text{ and } \lim_{d \to \infty} \phi(d) = 0.$$

The first condition imply that even if the new instance has the same input as one training data sample, we do not consider it to be 100% reliable, as the relation linking the input feature space and the output classes is not necessarily a function. From $\underline{P}_{(1),\boldsymbol{x}}, \ldots, \underline{P}_{(k),\boldsymbol{x}}$, we then obtain a combined lower prevision $\underline{P}$ such that

$$\underline{P}_{\boldsymbol{x}} = \frac{1}{k} \sum_{i=1}^{k} \underline{P}_{(i),\boldsymbol{x}}.$$

Using $\underline{P}_{\boldsymbol{x}}$ as the final uncertainty model for the true class of $\boldsymbol{x}$, one can predict its final class, either as a single class by using a maximin-like criteria or as a set of possible classes by using maximality or interval dominance. Using maximality or interval dominance is a good way to treat both ambiguity or large distances with the nearest neighbours. Indeed, if all nearest neighbours agree on the output class and are close to the new instance, the obtained lower prevision $\underline{P}_{\boldsymbol{x}}$ will be precise enough so that the criteria will end up pointing only one possible class (i.e., $\widehat{\Omega}_M, \widehat{\Omega}_I$ will be singletons). On the contrary, if nearest neighbours disagree or are far from the new instance, $\underline{P}_{\boldsymbol{x}}$ will be imprecise or indecisive, and $\widehat{\Omega}_M, \widehat{\Omega}_I$ will contain several possible classes.

3.1 Using lower previsions to choose k

A problem when using the k-nearest neighbour procedure is to choose the "best" number k of neighbours to consider. This number is often selected as the one achieving the best performance in a cross-validation procedure, but k-NN rules can display erratic performances if k is slightly increased or decreased, even if it is by one.

We propose here a new approach to guide the choice of k, using the features of lower previsions: we propose to choose the value k achieving the best compromise between imprecision and precision, estimated respectively from the number of optimal classes selected for each test sample, and from the percentage of times where the true class is inside the set of possible ones.

Let $(\boldsymbol{x}_{N+1}, y_{N+1}), \ldots, (\boldsymbol{x}_{N+T}, y_{N+T})$ be the test samples. Given a value k of nearest neighbours, let $\Omega^k_{M,i}$ denote the set of classes retrieved by maximality criteria for $\boldsymbol{x}_{N+i}$, and $\delta^k_i : 2^{|\mathcal{Y}|} \to \{0, 1\}$ the function such that $\delta^k_i = 1$ if $y_{N+i} \in \Omega^k_{M,i}$ and 0 otherwise. That is, δ^k_i is one if the right answer is in the set of possible classes. Then, we propose to estimate the informativeness Inf_k and the accuracy Acc_k of our k-NN method as:

$$Inf_k = 1 - \frac{\sum_{i=1}^{T} |\Omega^k_{M,i}| - T}{T(M-1)} \quad ; \quad Acc_k = \frac{\sum_{i=1}^{T} \delta^k_i}{T}$$

Note that informativeness has value one iff $|\Omega^k_{M,i}| = 1$ for $i = 1, \ldots, T$, that is decisions are precise, while accuracy measures the number of times the right class is in the set of possible classes. This means that the less informative is a classifier, the more accurate it will be, since the right answer will be in the set of possible classes every time. We then estimate the global performance GP_k as the value $GP_k = \beta Inf_k + (1-\beta)Acc_k$, that is a weighted average between precision and accuracy, with $\beta \in [0,1]$ the importance given to informativeness. Letting k vary, we then select the best value k^* as

$$k^* = arg \min_{k=1,\dots,N} GP_k.$$

The idea of this rule is to choose the value k^* achieving the best compromise between informativeness and accuracy (as some evaluation methods used for experts in classical probabilities).

3.2 Precise training samples and unitary costs

Let us now consider a particular case, namely the one where all training samples $\boldsymbol{x}_i$ have a single class y_i as output, and where the cost function (called here unitary) f_ω of choosing ω is $f_\omega(\omega') = 1 - \delta_{\omega,\omega'}$ where $\delta_{\omega,\omega'}$ is the classical Kronecker delta ($= 1$ if $\omega = \omega'$, zero otherwise). This assumptions corresponds to the one of classical k-NN procedures. Given these cost functions and a lower prevision $\underline{P}$ on $\mathcal{Y}$, the lower expectation for f_ω is

$$\underline{E}(f_\omega) = \underline{E}(\{\omega\}^c) = 1 - \overline{E}(\{\omega\}),$$

that is one minus the upper probability of the singleton ω. Similarly, the upper expectation of f_ω is one minus the lower probability of the singleton ω.

The lower prevision $\underline{P}_i$ and its natural extension $\underline{E}_i$ modeling our uncertainty about the output of a training sample $\boldsymbol{x}_i$ is simply, for any $f \in \mathcal{L}(\mathcal{Y})$, the value $\underline{E}_i(f) = f(y_i)$ where y_i is the output of $\boldsymbol{x}_i$. We also have $\underline{E}_i(f) = \overline{E}_i(f)$, and can now show that our method extends classical k-NN

Proposition 1. *Let k be the number of nearest neighbours considered. If training samples are precise, costs unitary and discounting rates $\epsilon_{(1)} = \ldots = \epsilon_{(k)} = \epsilon$, then the method used with a maximin decision criteria gives the same result as a classical k-NN rule.*

Proof. Let us consider a given $\omega \in \mathcal{Y}$ and its unitary cost function f_ω. Let us now compute the upper expectation of f_ω, or equivalently one minus the lower probability of $\{\omega\}$. Given the k nearest neighbour, the lower probability $\underline{E}(\{\omega\})$ of $\{\omega\}$ is

$$\underline{E}(\{\omega\}) = \frac{1}{k}\sum_{i=1}^{k} \epsilon\delta_{\omega,y_{(i)}} + (1-\epsilon)\inf f_\omega = \frac{\epsilon}{k}\sum_{i=1}^{k}\delta_{\omega,y_{(i)}}.$$

The highest value of $\underline{E}(\{\omega\})$ is reached for the value $\omega \in \mathcal{Y}$ which have the maximal number of representative in the k neighbours, and since the value maximising this lower probability is the same as the one minimising the upper expectation of unitary cost functions, this finishes the proof.

Proposition 2. *Let k be the number of nearest neighbours considered. If training samples are precise, costs unitary and discounting rates $\epsilon_{(i)} = w_i$ are equal to some weights, then the method used with a maximin decision criteria gives the same result as a weighted k-NN rule with the same weights.*

Proof. Similar to the proof of Prop. 1.

The case of precise training samples and unitary costs have another interesting property, namely the one that the set of possible classes obtained by maximality criteria coincide with the one obtained by interval dominance. This avoids any choice and allows using computational procedures used for interval-dominance, which are simpler.

Proposition 3. *Let k be the number of nearest neighbours considered. If training samples are precise and costs unitary, then $\widehat{\Omega}_M = \widehat{\Omega}_I$ for any new instance.*

Proof. To prove this proposition, we will simply show that for ω, ω', the two conditions to have $\omega \geq_I \omega'$ and $\omega \geq_M \omega'$ both coincide in this particular case. First, we have $\omega \geq_M \omega'$ if and only if $\underline{E}(\mathbf{1}_{(\{w\})} - \mathbf{1}_{(\{w'\})}) > 0$. Using Eq. and the particular case that we consider here, we have

$$\underline{E}(\mathbf{1}_{(\{w\})} - \mathbf{1}_{(\{w'\})}) = \frac{1}{k}\left(\sum_{i=1}^{k} \epsilon_{(i)}\delta_{\omega,y_{(i)}} - \sum_{i=1}^{k} \epsilon_{(i)}\delta_{\omega',y_{(i)}} - \sum_{i=1}^{k}(1-\epsilon_{(i)})\right).$$

The last part of the equation right-hand side being due to the fact that $\inf_{\omega \in \mathcal{Y}}(\mathbf{1}_{(\{w\})} - \mathbf{1}_{(\{w'\})}) = -1$ if $\omega \neq \omega'$. Hence, $\omega \geq_M \omega'$ iff the number between parenthesis is positive. Now, we have that $\omega \geq_I \omega'$ if and only if $\underline{E}(\mathbf{1}_{(\{w\})}) \geq \overline{E}(\mathbf{1}_{(\{w'\})})$. In our particular case, this becomes

$$\frac{1}{k}\sum_{i=1}^{k} \epsilon_{(i)}\delta_{\omega,y_{(i)}} \geq \frac{1}{k}\left(\sum_{i=1}^{k} \epsilon_{(i)}\delta_{\omega',y_{(i)}} + \sum_{i=1}^{k}(1-\epsilon_{(i)})\right).$$

Moving the right hand side to the left finishes the proof.

4 Experiments

Since Proposition 2 indicates that the results of the proposed method can be made equivalent (in terms of prediction accuracy) to those of a weighted k-NN method, we refer to studies comparing the results of different weighted k-NN method to have an idea about the accuracy of the method.

Instead, we have preferred to experiment our method to select the best number k of nearest neighbours on some classical benchmark problems. We used a leave-one-out validation method. The class of each sample is predicted using the $N-1$ remaining samples. Inf_k, Acc_k and GP_k are averaged over the N obtained results. We also computed the average error rate using a maximin criterion, which gives results equivalent to the weighted k-NN with weights given by the discounting factor.

Name	# instances	# input variables	# output classes
Glass	214	9	6
Image segmentation	2100	19	7
Ionosphere	351	9	2
Letter recognition	2500	16	26

Table 1. Experiment data sets

As discussing and optimising ϕ is not the topic of the paper, we consider the simple heuristic where, for a given training data $(\boldsymbol{x}, y)$, $\phi(d_{\boldsymbol{x}}) = \exp^{-d/\overline{d_y}}$, with $\overline{d_y}$ the average distance between elements of the training set having y for class. We fix $\epsilon_0 = 0.99$, in order to not increase too quickly the imprecision.

Four different classification problems taken from the UCI repository [12] are considered. They are summarized in Table 1 . Results obtained for each of them are summarized in Fig 1. In each graphs are displayed, for different values of k nearest neighbours, the informativeness Inf_k, the precision Acc_k, the global score GP_k as well as the precision obtained by using a maximin criterion, equivalent to the one obtained with a weighted k-NN method using the discounting weights.

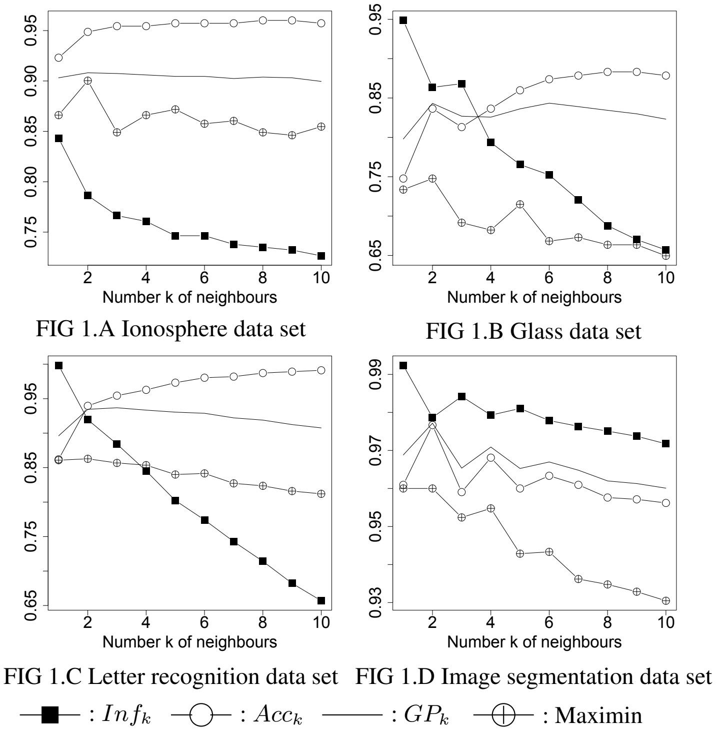

FIG 1.A Ionosphere data set

FIG 1.B Glass data set

FIG 1.C Letter recognition data set

FIG 1.D Image segmentation data set

Fig. 1. Experiment results

Note that, here, both the choices of β, of ϵ_0 and of $\phi()$ are of importance, for they will directly influence the imprecision of $\underline{P}_{\boldsymbol{x}}$ and hence the decision imprecision concerning the class of $\boldsymbol{x}$ and the optimal k^*. As could be expected, the informativeness globally decreases with the number k of nearest neighbours, while the number of sample x_i whose true class is in the set of optimal classes $|\Omega^k_{M,i}|$ globally increases. Note that this imprecision is due to two different causes: the presence of conflicting information in (in this case, the different classes to which belongs the neighbours are optimal) and distance of the neighbours to the sample (in this case, $\underline{P}_{\boldsymbol{x}}$ is very imprecise and no class dominates another, i.e., they are all optimal).

The increase in informativeness that we can see when going from $k = 2$ to $k = 3$ for the Glass and Image segmentation data sets are due to the fact that immediate neighbours provide conflicting information that do not make decisions less informative, but provoke, for some sample, a decision shift from their true class to a false class. Such an increase is then the clue that some classes boundaries may be quite difficult to identify in the input space. A smooth decrease of informativeness is then the clue that there are no significant conflict in the information provided by neighbours, as for the ionosphere and letter recognition data.

The initial number of samples that have imprecise classifications due to the distance with their neighbours can be evaluated from the informativeness for $k = 1$. Indeed, if $k = 1$, there can be no conflict between neighbours, and the imprecise classification can only come from the large distance and the resulting discounting weight. It is therefore also a good way to evaluate the density of the data set, and its representativeness (for example, points in the ionosphere data set seems to have large distances between them, compared to the others).

Although they could probably be improved by optimised choices of the metric, of parameters β, ϵ_0, $\phi()$, our results show that allowing for a small imprecision can improve significantly the resulting classification, and the confidence we have in the classifier answer, without adding additional parameters such as a rejection or distance threshold. They also indicate that, in general, best results are obtained for a small number of neighbours. Finally, if one wants a unique class as answer, it is always possible to come back to the solution of a classical weighted k-NN method. An alternative would be to use another classifier and its answer to precisiate the imprecise answer given by our method.

5 Conclusion and perspectives

In this paper, we have defined a first K-NN method based on lower previsions (equivalent to convex probability sets). As lower previsions are very generic models of uncertainty, using them allows to handle labels coming from expert opinions expressed in very different ways. Using the theory of lower previsions also allows to settle the problem of ambiguity (conflicting information) and absence of neighbours close to a given instance, without adding additional parameters. This can be done by using decision rules that selects sets of possible (i.e., optimal) classes rather than single ones when information delivered by neighbours is ambiguous or unreliable.

Using this particular feature of lower previsions, we have proposed a simple and new means to select the "best" number k of nearest neighbours to consider. Namely, the number that achieves the best balance between accuracy (good classification) and precision (decision retaining only a small number of classes).

This paper have exposed the basics of a K-NN method using lower previsions. Many surrounding topics remains to be investigated, among which:

- how to distinguish imprecise decisions due to ambiguity from those due to unreliable (i.e. "far away") neighbours ?
- how to optimise (as done in [13]) the whole procedure so that it can give better results for a given problem ?
- how the framework of lower previsions can help in solving the problem of instances having uncertain / missing input values ?
- how does this method compare to other (basic) classification methods using lower previsions, such as the Naive credal classifier [14] ?

References

1. Fix, E., Hodges, J.: Discriminatory analysis, nonparametric discrimination: consistency properties. Technical Report 4, USAF School of Aviation Medicine (1951)
2. Dudani, S.: The distance-weighted k-nearest neighbor rule. IEEE Trans. Syst. Man. Cybern. **6**(325-327) (1976)
3. Dubuisson, B., Masson, M.: A statistical decision rule with incomplete knowledge about classes. Pattern Recognition **26** (1993) 155–165
4. Denoeux, T.: A k-nearest neighbor classification rule based on dempster-shafer theory. IEEE Trans. Syst. Man. Cybern. **25** (1995) 804–813
5. Hüllermeier, E.: Case-based approximate reasoning. Volume 44 of Theory and decision library. Springer (2007)
6. Walley, P.: Statistical reasoning with imprecise Probabilities. Chapman and Hall, New York (1991)
7. Miranda, E.: A survey of the theory of coherent lower previsions. Int. J. of Approximate Reasoning **48** (2008) 628–658
8. Dubois, D., Prade, H.: Possibility Theory: An Approach to Computerized Processing of Uncertainty. Plenum Press, New York (1988)
9. Shafer, G.: A mathematical Theory of Evidence. Princeton University Press, New Jersey (1976)
10. Walley, P.: The elicitation and aggregation of beliefs. Technical report, University of Warwick (1982)
11. Troffaes, M.: Decision making under uncertainty using imprecise probabilities. Int. J. of Approximate Reasoning **45** (2007) 17–29
12. Asuncion, A., Newman, D.: UCI machine learning repository [http://www.ics.uci.edu/∼mlearn/MLRepository.html] (2007)
13. Zouhal, L., Denoeux, T.: An evidence-theoretic k-nn rule with parameter optimization. IEEE Trans. on Syst., Man, and Cybern. **28** (1998) 263–271
14. Zaffalon, M.: The naive credal classifier. J. Probabilistic Planning and Inference **105** (2002) 105–122