Pattern Classification through Fuzzy Likelihood

Giovanni Gallo, Rosa Maria Pidatella, Masoumeh Zeinali

University of Catania, Italy
University of Catania, Italy
University of Tabriz, Iran
gallo@dmi.unict.it, rosa@dmi.unict.it, mzds_21@yahoo.com

Keywords. Fuzzy; Bayes rule; classification; likelihood estimation

1 Introduction

Generalization of classification rules is a fundamental issue in automatic pattern recognition. Overfitting a classifier on the training data is a well known problem and it has been the focus of a lot of research in the recent decade. Fuzzy techniques naturally provide soft representation of functions that could be adapted to address some of the overfitting/generalization dilemma.
In the literature there are a lot of papers concerning fuzzy theory as a mean for classifying and extracting information from a huge amount of data in a human-like fashion. Many authors have studied how to obtain a membership function of a fuzzy set by ad hoc heuristics, histograms, nearest-neighbor, etc. In [1] a definition of fuzzy likelihood measure was proposed in the similarity estimation context, while [2] puts the basis of adaptive fuzzy likelihood algorithms in the context of system theory and fuzzy logic.
In this paper we propose a new approach to supervised classification based on a novel proposal for a fuzzy likelihood function. This new function leads to a fuzzy version of Bayes Rules for Maximum a Posteriori classification(MAP). The performances of the proposed new method are close to the performances of classical methods and the new technique provides several advantages. Classification can be done using a confidence threshold set by the user; moreover an automatic criterion to signal cases when classification cannot be safely done is intrinsecally provided by our approach.
Starting from the histograms of the observed data, we provide a simple way to obtain the membership function of a fuzzy set approximating the data distribution. This is obtained combining together the raw data histograms with their successively smoothed versions. A posterior probability is, in turn, obtained through a suitable fuzzy version of the Bayesian formula. It is important to note that, since our likelihoods are fuzzy numbers, a careful translation in terms of *restricted fuzzy arithmetic* has to be done for the classical Bayes rule in order to obtain meaningful probabilities.
To classify a member in a set we adopt the *overtaking* relation between fuzzy numbers introduced in [3]. The overtaking mimics an ordering relation between

fuzzy numbers that depends on an assigned threshold value. The ordering imposed by the overtaking relation translates immediately into a dominance of the posterior probability of a class over another for a given observed value. In this way a crisp classification is eventually obtained. The proposed method has been tested on some standard data sets and the results are reported below.
The authors have implemented the proposed ideas in Matlab and performed classification over some standard benchmarks. In all cases the results have been close to the theoretical optimal error rate.
The rest of this paper is organized as follows: Section 2 describes our fuzzyfication procedure to obtain a fuzzy version for likelyhood distribution from a given training set; Section 3 introduces a fuzzy version of Bayes rule and explains how to wisely use the arithmetic of fuzzy number to keep the results of computation within reasonable bounds; Section 4 recalls the concept of overtaking, a possible pseudo-ordering for fuzzy numbers; Section 5 reports of some of the experimental test that have been performed on some benchmark data sets.

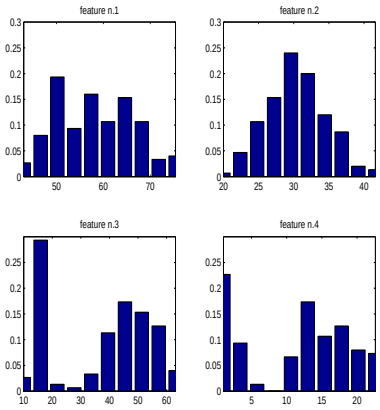

Fig. 1. Histograms of Fisher's irises data set for the four features.

2 Histogram fuzzification

In this section we show how to construct fuzzy likelihoods directly from the data by using a membership construction algorithm. Our technique applies to one dimensional labelled data set. The data are a set of pairs (x, l) where x is the measure of an observed feature, i.e. it is a crisp number, and l is the indicator of a class and ranges over a finite set L of labels.
Let $[x_{min}, x_{max}]$ the range of the observed data. We choose to partition it into h suitable number of equally spaced discrete bins (i.e. uniform quantization). The

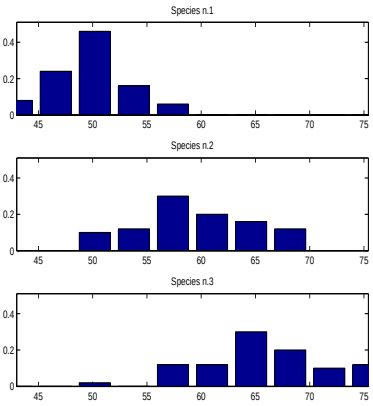

Fig. 2. Histograms of the three species of flowers for the first feature.

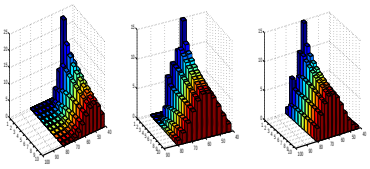

Fig. 3. Iterated convolution of the histograms of figure 2

relative frequencies of the data in the bins form the standard crisp histogram approximating the training data. In practice, if f_i is the relative frequence of data falling in the i-th bin, the histogram is a vector $(f_1, ..., f_h)$. In figure 1 we show the histograms of the whole population of the classical Fisher's irises data set for the four registered features of the flowers. In figure 2 we show the separate histograms of the three species of flowers for the first feature. We are interested into assigning a fuzzy membership function $\overline{m}$ to the bins i.e. to assign a fuzzy number $\overline{m}(i)$ to each bin. This function is indeed our proposed fuzzy likelihood. In this paper we choose a computational representation of a fuzzy number as a finite sequence of nested intervals $(a, b)[\alpha]$. According to fuzzy arithmetic jargon each interval corresponds to successive α values

$$1 \geq \alpha_1 \geq \cdots \geq \alpha_l \geq 0.$$

In our proposal $\alpha_1 = 1$ and the first α-cut of $\overline{m}(i)$ is the singleton $\{f_i\}_{i=1,...,h}$. To obtain the successive α-cut we perform the convolution of $(f_1, ..., f_h)$ with a suitable smoothing unitary kernel $K = (\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$.
Let $({f_1}^0, ..., {f_h}^0) = (f_1, ..., f_h)$. We define :

$$({f_1}^{(i)}, ..., {f_h}^{(i)}) = ({f_1}^{(i-1)}, ..., {f_h}^{(i-1)}) * K$$

In figure 3 we show the results obtained with the iterated convolution of the original histogram of figure 2 for iris data. The i-th α-cut for the j-th bin $({a_j}^{(i)}, {b_j}^{(i)})$ is obtained as follows:
${a_j}^{(i)} = min({f_j}^{(0)}, ..., {f_j}^{(i)}), \quad {b_j}^{(i)} = max({f_j}^{(0)}, ..., {f_j}^{(i)})$.
For sake of simplicity we write: $(a, b)[\alpha_i] \equiv ({a_j}^{(i)}, {b_j}^{(i)})$.

In figure 4 we show the calculated membership function obtained with the kernel: $K = (\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$ over the first feature of the three species of the iris data set.

3 Fuzzy Bayes rule

Following [4], in this section we illustrate the use of the *restricted arithmetics* which is necessary to adopt when using Bayes rule in order to obtain values within the range $[0, 1]$ so that they can be soundly considered as posterior probabilities.
Suppose we have a finite set $X_n = \{x_1, \ldots, x_n\}$ and let P the probability function of each x_i such that:

$$P(\{x_i\}) = p_i, \qquad i = 1, \ldots, n, \qquad 0 < p_i < 1, \qquad \sum_{i=1}^{n} p_i = 1$$

Then P is a discrete probability function on X_n.
If one or more p_i are uncertain, we can substitute p_i with $\bar{p_i}$, a fuzzy number such that each α-cut of $\bar{p_i}$ is contained within $[0, 1]$. With abuse of notation we write:

$$\bar{P}(\{x_i\}) = \bar{p}_i, \qquad 0 < \bar{p}_i < 1, \qquad i = 1, \ldots, n$$

Let's indicate the α-cut of the fuzzy number $\bar{p}_i$ with $\bar{p}_i[\alpha]$. We can choose $p_i \in \bar{p}_i[\alpha]$ if and only if we satisfy the condition: $\sum_{i=1}^{n} p_i = 1$ for every $\alpha \in [0,1]$. With this hypothesis, we can define now the fuzzy conditional probability. Let $X_k = \{x_1, \ldots, x_k\} \subseteq X_n$ with $1 \leq k < n$. Then:

$$\bar{P}(X_k)[\alpha] = \left\{ \sum_{i=1}^{k} p_i \mid S \right\}$$

where S means the statement:

$$S = \ p_i \in \bar{p}_i[\alpha], \ \ i = 1, \ldots, n, \ \sum_{i=1}^{n} p_i = 1$$

In [4] it is proven that $\bar{P}(X_k)[\alpha]$ is the α-cut of the fuzzy probability $\bar{P}(X_k)$. Let $X_{1k} = \{x_1, \ldots, x_k\}$, $X_{lm} = \{x_l, \ldots, x_m\}$, $1 \leq l \leq k \leq m \leq n$ be two not disjoint subsets of X_n. As in [4], we define the fuzzy conditional probability of X_{1k} given X_{lm} as

$$\bar{P}(X_{1k} \setminus X_{lm}) = \left\{ \frac{\sum_{i=l}^{k} p_i}{\sum_{j=l}^{m} p_j} \mid S \right\}$$

where S is the same above statement.
To better illustrate the ideas reported above, let's turn to the iris data set. Let $\bar{P}(C_j \setminus S_q)$ be the fuzzy likelihood of the species $S_q, q = 1, 2, 3$ with the characteristic $C_j, j = 1, \ldots, h$ and let $\bar{P}(S_q \setminus C_j)$ be the fuzzy posteriori probability of the species S_q with the characteristic C_j . We apply the Bayes rule, using the restricted arithmetics, in order to obtain values for the probability within the range $[0, 1]$ for the posterior probability:

$$\bar{P}(S_q \setminus C_j) = \frac{\bar{P}(C_j \setminus S_q)}{\sum_{k=1}^{3} \bar{P}(C_j \setminus S_k)} \tag{1}$$

To apply restricted arithmetics it is useful to investigate the functional behaviour of the terms in (1).
We put, for simplicity:

$$p_{qj} = \bar{P}(C_j \setminus S_q), \quad q = 1, 2, 3 \quad j = 1, \ldots, h$$

Let us study the behaviour of the functions:

$$f_q(p_{1j}, p_{2j}, p_{3j}) = \frac{p_{qj}}{p_{1j} + p_{2j} + p_{3j}}, \quad q = 1, 2, 3 \quad j = 1, \ldots, h.$$

For sake of simplicity let fix $q = 1$.
Observe that:

$$\frac{\partial f_1}{\partial p_{1j}} > 0, \qquad \frac{\partial f_1}{\partial p_{2j}} < 0, \qquad \frac{\partial f_1}{\partial p_{3j}} < 0.$$

We then obtain:

$$min\, f_1 = f_1(min\, p_{1j}, max\, p_{2j}, max\, p_{3j})$$

$$max\, f_1 = f_1(max\, p_{1j}, min\, p_{2j}, min\, p_{3j})$$

then: $\bar{P}(S_1 \setminus C_j)[\alpha] = [min\, f_1, max\, f_1]$.
The same derivation can be carried on for $q = 2, 3$.

4 Overtaking

There are many ways to compare fuzzy numbers [5]. In [3], an *overtaking* operator is introduced first on intervals and then it is generalized to fuzzy numbers. For sake of self-containment the construction introduced in [3] is here reported. First let us define a function $\sigma(A, B)$ for pairs of intervals A and B. Let us first assume that neither A or B are reduced to a crisp number. Observe that if A_l, A_u, B_l, B_u are, respectively, the lower and upper bounds of intervals A and B, only the cases reported in the following table are possible.

$A_u \leq B_u$	$A_u \leq B_l$	$A_l \leq B_u$	$A_l \leq B_l$	$\sigma(A, B)$
T	T	T	T	0
T	F	T	T	$\frac{A^u - B^l}{w(A)}$
T	F	T	F	1
F	F	T	T	$\frac{B^u - B^l}{w(A)}$
F	F	T	F	$\frac{B^u - A^l}{w(A)}$
F	F	F	F	1

Table 1. σ values for different positions of intervals A and B

where *w(A)* is the width of the interval A.
Now let us consider some special cases to be treated separately. They are:
(i) $A_l = A_u = a$, i.e. interval A is degenerate in a single point a, but $B_l < B_u$;
(ii) $A_l < A_u$ but $B_l = B_u = b$, i.e. interval B is degenerate in a single point b;
(iii) both A and B are degenerate intervals.
In case (i)

$$\sigma(A, B) = \begin{cases} 0 & \text{if } a \leq B_l \\ 1 & \text{if } a > B_l \end{cases} \tag{2}$$

In case (ii)

$$\sigma(A, B) = \begin{cases} 1 & \text{if } b \leq A_u \\ 0 & \text{if } b > A_u \end{cases} \tag{3}$$

In case (iii)

$$\sigma(A, B) = \begin{cases} 0 & \text{if } a < b \\ 1 & \text{if } a = b \\ 1 & \text{if } a > b \end{cases} \tag{4}$$

The δ-overtaking operator is defined as follows. Given two intervals A, B and a real number $\delta \in [0, 1]$, A overtakes B if $\sigma(A, B) \geq \delta$ or:

$$A \geq_\delta B \iff \sigma(A, B) \geq \delta$$

The overtaking depends then on the choosen δ value.
The extension of the δ-overtaking relation to pairs of fuzzy numbers, once these are defined using α-cuts, is as follows. Let us assume that fuzzy numbers A and B are defined as two finite and equal sized collections of α-cuts

$$A = \{A[\alpha_i]\}, \quad B = \{B[\alpha_i]\}$$

$$0 \leq \alpha_k \leq \alpha_{k-1} \leq \cdots \leq \alpha_1 \leq 1.$$

The degree of overtaking of A and B is

$$overtaking(A, B) = \sum_{i=1}^{k} w_i \cdot \sigma(A[\alpha_i], B[\alpha_i])$$

where $w_1, w_2, \ldots w_k \in [0, 1]$ and $\sum_{i=1}^{k} w_i = 1$.
We say that A δ-overtakes B if

$$overtaking(A, B) > \delta.$$

5 Experiments

To verify the performance of the proposed classification technique we made several experiments on public data sets commonly used by the Pattern Recognition community as benchmarks. In this section we report the experiment protocol and the results obtained, drawing some conclusive remarks about the proposed technique.
The data sets used in the experiments are:

1. Fisher's irises data set (150 records, 4 features, 3 classes);
2. Diabetes Pima Indians data set (768 records, 8 features, 2 classes);
3. Italian wines quality data set (178 records, 13 features, 3 classes).

All three data sets may be retrieved from the public data repository at the URL [6].
Observe that the data sets taken as benchmarks are intrinsically multidimensional: it is well known that, achieving good classification results for them requires to jointly consider all of their features. On the other hand at this stage of our investigation the proposed fuzzy Bayesian classification algorithm has been developed only to process single featured data. Research to generalize it to the multi-featured case is ongoing.
Only one feature at the time has been hence considered, performing 25 experiments in total. The results have not be evaluated in absolute terms but in comparison between the performance of classical Bayes MAP classifier and the proposed fuzzy generalization. We investigated mainly the discriminative power of the proposed algorithm.

The experimental scheme that has been carried out is as follows. The experiments are focused to estimate the training error that can be achieved with the fuzzy Bayes rule in comparison with the crisp one. In particular, we checked if the percentage of hits, i.e. correctly classified records of the training set obtained with fuzzy method, is greater than the percentage obtained with the crisp case. Similarly we are interested in checking if the percentage of misses, i.e. uncorrectly classified records, decreases in comparison with the percentage obtained with the crisp case.
It turns out that the fuzzy approach reduces the misses but at the price of labelling some records as unclassified. This is indeed a point of our approach: the proposed algorithm automatically weights the evidence leading to classification. When evidence is not sufficiently strong, instead of risking a wrong labelling, it declares the record as unclassified. In real applications this *problematic* datum could be hence passed to a more sophisticated and typically more costly classifier. A summary of this comparison is presented in table 2.

Hits rate	**Misses rate**	**Unclassified rate**	$\delta = \mathbf{0.25}$	$\delta = \mathbf{0.50}$	$\delta = \mathbf{0.75}$
less	more	many	0.00	0.00	0.00
less	more	few	0.04	0.08	0.07
less	less	many	0.13	0.03	0.21
less	less	few	0.29	0.17	0.29
more	more	many	0.00	0.00	0.00
more	more	few	0.00	0.00	0.00
more	less	many	0.00	0.00	0.04
more	less	few	0.54	0.72	0.39

Table 2. Test percentage of hits, miss and unclassified for feature 1, training error

Table 2 shows that the best performances are obtained choosing $\delta = 0.5$ for the overtaking relation, although good results are evident with a lower δ value. About a three fourth of our classification experiments resulted into an increased hits rate, a decreased misses rate and a moderate number of unclassified records (see last line of table 2). In about another fourth of experiments the misses rate is reduced, but the somehow prudential policy of our algorithm keeps in the low the percentage of hits (see the fourth line of table 2).

6 Conclusions

This paper has introduced a possible generalization into a fuzzy framework of the classical MAP classifier. The new algorithm is based on histogram smoothing and on fuzzy version of Bayes rule. Experimental results have shown that the algorithm could be useful in practical pattern recognition providing both a good classifier and an automatic sieve for ambiguous data to be treated with more complex techniques.

Although only one feature case has been reported here, research is ongoing to apply the same ideas to the multidimensional case.

References

[1] Ding S., Jin F. (2008) A novel fuzzy likelihood measure algorithm *Intern. Conf. on Computer Science and Software Engineering*, pp. 945–948.

[2] Osoba O., Mitaim S., Kosko B. (2011) Bayesian Inference with Adaptive Fuzzy Priors and Likelihoods *IEEE Trans on Systems, Man and Cybernetics*, Vol. 41, n. 5, pp. 1183–1197.

[3] Anile A. M., Spinella S.(2004) Modeling Uncertain Sparse Data with Fuzzy B-splines *Reliable Computing*, Vol. 10, n. 5, pp. 335–355.

[4] Buckley J. J. (20..) Fuzzy Probabilities *Studies in Fuziness and Soft Computing* Springer ed.

[5] Dubois D., Kerr E., Mesiar R., Prade H., (2001) Fuzzy Interval Analysis in Fundamentals of Fuzzy Sets, Dubois D., Prade H. eds, The Handbook of Fuzzy Sets, pp. 483-581, Kluwer.

[6] URL: http://archive.ics.uci.edu/ml/