156

Estimations of the Bayes Classifier Error with Fuzzy Observations

Abstract. The paper presents the problem estimation of the error in the Bayes classifier. The model of pattern recognition with fuzzy observations and the zero-one loss function was assumed. Deterioration of the quality of the classification in the relation to the task with crisp observations was showed for this model. The exact computation of the error concerns a full probabilistic and fuzzy information. Received results were compared from the bound on the probability of error based on information energy for fuzzy events. Numerical example presents that this bound is worse than introduced in the work.

1 Introduction

Many paper present aspect of fuzzy and imprecise information in pattern recognition [3], [4], [14], [15], [16]. In [10] formulated the pattern recognition problem with fuzzy classes and fuzzy information and consider the following tree situations:

- fuzzy classes and exact information,
- exact classes and fuzzy information,
- fuzzy classes and fuzzy information.

The classification error is the ultimate measure of the performance of a classifier. Competing classifiers can also be evaluated based on their error probabilities. Several studies have previously described the Bayes probability of error for a single-stage classifier [1], [2], for a combining classifiers [17] and for a hierarchical classifier [7], [9]. Some studies pertaining to bounds on the probability of error in fuzzy concept are given in [4], [10], [13], [11].

In this paper, our aim is to obtain the error probability when the decision in the classifier is made according to the Bayes method. In our study we consider the situation with exact classes and fuzzy information on object features, i.e. when observations of the features are represented by the fuzzy sets. The received results are compared with the bound on the probability of error based on information energy.

The contents of the work are as follows. Section 2 introduces the necessary background and describes the Bayes classifier. In section 3 the basic notions of fuzzy theory are presented. In section 4 we presented the difference between the probability of misclassification for the fuzzy and crisp data in Bayes optimal classifier.

2 Bayes classifier

Bayesian decision theory is a fundamental statistical approach to the problem of pattern classification. This approach is based on quantifying the tradeoffs between various classification decision using probability and the costs that accompany such decision. It makes the assumption that the decision problem is posed in probabilistic terms, and that all of the probability values are known.

A pattern is represented by a set of d features, or attributes, viewed as a d-dimensional feature vector $x \in \Re^d$.

Let us consider a pattern recognition problem, in which the class label ω is a random variable taking values in the set of class labels $\Omega = \{\omega_1, ..., \omega_c\}$. The *priori probabilities*, $P(\omega_i), i = 1, ..., c$ constitute the probability mass function of the variable ω, $\sum_{i=1}^{c} P(\omega_i) = 1$. Assume that the objects from class ω_i are distributed in $x \in \Re^d$ according to the *class-conditional probability density function* $p(x|\omega_i)$, $p(x|\omega_i) \geq 0$, $\forall x \in \Re^d$, and $\int_{\Re^d} p(x|\omega_i)dx = 1, i = 1, ..., c$.

Given the prior probabilities and the *class-conditional probability density functions* we can calculate the *posterior probability* that the true class label of the measured x is ω_i using the Bayes formula

$$P(\omega_i|x) = \frac{P(\omega_i)p(x|\omega_i)}{p(x)} \tag{1}$$

where $p(x) = \sum_{i=1}^{c} P(\omega_i)p(x|\omega_i)$ is the unconditional likelihood of $x \in \Re^d$.

Equation (1) gives the probability mass function of the class label variable ω for the observed x. The decision for that particular x should be made with respect to the posterior probability.

The "optimal" Bayes decision rule for minimizing the risk (expected value of the loss function) can be stated as follows: Assign input pattern x to class ω_i for which the conditional risk

$$R^*(\omega_i|x) = \sum_{j=1}^{c} L(\omega_i, \omega_j)P(\omega_j|x) \tag{2}$$

is minimum, where $L(\omega_i, \omega_j)$ is the loss incurred in deciding ω_i when the true class is ω_j. The Bayes risk, denoted R^*, is the best performance that can be achieved. In the case of the zero-one loss function

$$L(\omega_i, \omega_j) = \begin{cases} 0, \, i = j \\ 1, \, i \neq j \end{cases},$$

the conditional risk becomes the conditional probability of misclassification and optimal Bayes decision rule is as follows:

$$R^*(\omega_i|x) = \max_i P(\omega_i|x). \tag{3}$$

Let Ψ^* be a classifier that always assigns the class label with the largest posterior probability. The classifier based on Bayes rule is the following:

$$\Psi^*(x) = \omega_i \quad \text{if} \quad \omega_i = \arg\max_i P(\omega_i)p(x|\omega_i). \tag{4}$$

because the unconditional likelihood $p(x) = \sum_{i=1}^{c} P(\omega_i)p(x|\omega_i)$ is even for every class ω_i

3 Basic notions of fuzzy theory

Fuzzy number A is a fuzzy set defined on the set of real numbers $\mathbb{R}$ characterized by means of a membership function $\mu_A(x)$, $\mu_A : \mathbb{R} \to [0, 1]$:

$$\mu_A(x) = \begin{cases} 0 & \text{for } x \leq a, \\ f_A(x) & \text{for } a \leq x \leq c, \\ 1 & \text{for } c \leq x \leq d, \\ g_A(x) & \text{for } d \leq x \leq b, \\ 0 & \text{for } x \geq b, \end{cases}$$

where f_A and g_A are continuous functions, f_A is increasing (from 0 to 1), g_A is decreasing (from 1 to 0). In special cases it may be $a = -\infty$ and (or) $b = +\infty$. In this study, the special kinds of fuzzy numbers including triangular fuzzy numbers is employed. A triangular fuzzy numbers can be defined by a triplet $A = (a_1, a_2, a_3)$. The membership function is

$$\mu_A(x) = \begin{cases} 0 & \text{for } x \leq a_1, \\ (x - a_1)/(a_2 - a_1) & \text{for } a_1 \leq x \leq a_2, \\ (a_3 - x)/(a_3 - a_2) & \text{for } a_2 \leq x \leq a_3, \\ 0 & \text{for } x \geq a_3. \end{cases}$$

The width w_A of the fuzzy number A is defined as following value [5]:

$$w_A = \int_{-\infty}^{+\infty} \mu_A(x)dx. \tag{5}$$

A fuzzy information $\mathcal{A}_k \in \Re^d$, $k = 1, ..., d$ (d is the dimension of the feature vector) is a set of fuzzy events $\mathcal{A}_k = \{A_k^1, A_k^2, ..., A_k^{n_k}\}$ characterized by membership functions

$$\mathcal{A}_k = \{\mu_{A_k^1}(x_k), \mu_{A_k^2}(x_k), ..., \mu_{A_k^{n_k}}(x_k)\}. \tag{6}$$

The value of index n_k defines the possible number of fuzzy events for x_k (for the k-th dimension of feature vector). In addition, assume that for each observation subspace x_k the set of all available fuzzy observations (6) satisfies the orthogonality constraint [14]:

$$\sum_{l=1}^{n_k} \mu_{A_k^l}(x_k) = 1. \tag{7}$$

The probability of fuzzy event assume in Zadeh's form [18]:

$$P(A) = \int_{\Re^d} \mu_A(x) f(x) dx. \tag{8}$$

The probability $P(A)$ of a fuzzy event A defined by (8) represents a crisp number in the interval $[0, 1]$.

4 Estimations of the Bayes classifier error

4.1 Estimation of the Bayes classifier error with crisp observations

The error of Ψ^* is the smallest possible error, called the Bayes error. The overall probability of error of Ψ^* is the sum of the errors of the individual xs weighted by their likelihood values $p(x)$,

$$Pe(\Psi^*) = \int_{\Re^d} [1 - P(\omega_i^*|x)] p(x) dx. \tag{9}$$

It is convenient to split the integral into c integrals, one on each classification region. For this case class ω_i^* will be specified by the regions label. Then

$$Pe(\Psi^*) = \sum_{i=1}^{c} \int_{\Re_i^*} [1 - P(\omega_i|x)] p(x) dx \tag{10}$$

where $\Re_i^*$ is the classification region for class ω_i, $\Re_i^* \cap \Re_j^* = 0$ for any $i \neq j$ and $\bigcup_{i=1}^{c} \Re_i^* = \Re^d$. Substituting (1) into (10) we have [8]:

$$Pe(\Psi^*) = 1 - \sum_{i=1}^{c} \int_{\Re_i^*} P(\omega_i) p(x|\omega_i) dx. \tag{11}$$

In Fig. 1 the Bayes error is presented for the simple case of $x \in \Re$, $\Omega = \{\omega_1, \omega_2\}$ and $P(\omega_1|x) = 1 - P(\omega_2|x)$. According to (10) the Bayes error is the area under $P(\omega_2)p(x|\omega_2)$ in $\Re_1^*$ plus the area under $P(\omega_1)p(x|\omega_1)$ in $\Re_2^*$. The total area corresponding to the Bayes error is marked in black.

4.2 Estimation of the Bayes classifier error with fuzzy observations

When we have non-fuzzy observation of object features in Bayes classifier then recognition algorithm for zero-one loss function is given by (3) and probability

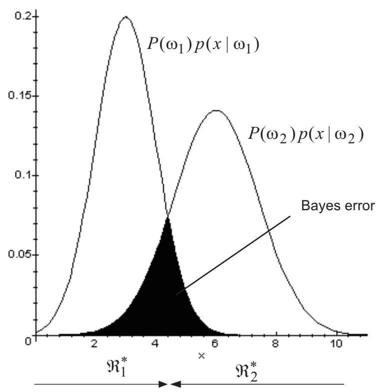

Fig. 1. The probability of error for Bayes optimal classifier when object features are non-fuzzy

of error is given by (10). Similarly, if (7) holds and we use probability of fuzzy event given by (8) the Bayes recognition algorithm for fuzzy observations $\tilde{A}$ is the following:

$$\Psi_F^*(\tilde{A}) = \omega_i \quad \text{if} \tag{12}$$

$$\omega_i = \arg\max_i P(\omega_i) \int_{\Re^d} \mu_{\tilde{A}}(x) p(x|\omega_i).$$

The probability of error $Pe(\Psi_F^*)$ for fuzzy data is the following:

$$Pe(\Psi_F^*) = 1 - \sum_{i=1}^{c} \sum_{\tilde{A} \in i} P(\omega_i) \int_{\Re_i^*} \mu_{\tilde{A}}(x) p(x|\omega_i) dx, \tag{13}$$

where $\tilde{A} \in i$ denote the fuzzy observations belongs to the i-th classification region.

When we use fuzzy information on object features instead of exact information we deteriorate the classification accuracy. The difference between the probability of misclassification for the fuzzy $Pe(\Psi_F^*)$ and crisp data $Pe(\Psi^*)$ in Bayes optimal classifier is the following:

$$Pe(\Psi_F^*) - Pe(\Psi^*) = \tag{14}$$

$$= \sum_{\tilde{A} \in \Re^d} \left(\int_{\Re^d} \mu_{\tilde{A}}(x) \max_i \{P(\omega_i) p(x|\omega_i)\} dx - \right.$$

$$\left. - \max_i \left\{ \int_{\Re^d} \mu_{\tilde{A}}(x) P(\omega_i) p(x|\omega_i) dx \right\} \right).$$

similarly as in [4].

The element of $\sum\limits_{\tilde{A}\in\Re^d}$ equals to 0 if and only if, for the support of fuzzy observation $\tilde{A}$, one of the i discriminant functions $[P(\omega_1)p(x|\omega_1),\ldots,P(\omega_i)p(x|\omega_i)]$ is uniformly larger than the others. Another interpretation is that the value of equation (14) depends only from these the observation, in whose supports intersect the discriminant functions.

4.3 Error bounds in terms of information energy for fuzzy observations

Some studies pertaining to bound on the probability of error in fuzzy concepts are presented in [13], [11]. They are based on information energy for fuzzy events. The information energy contained in the fuzzy event A is defined by [12]:

$$W(A) = P(A)^2 + P(\bar{A})^2, \tag{15}$$

where $P(\bar{A})$ is the complement set of A.

The information energy contained in the fuzzy information $\mathcal{A}$ is defined by [12]:

$$W(\mathcal{A}) = \sum_{l=1}^{k} P(A_l)^2. \tag{16}$$

The marginal probability distribution on fuzzy information $\mathcal{A}$ of the fuzzy event A is given by:

$$P_m(A) = \int\limits_{\Re^d} \mu_A(x)p(x)dx, \tag{17}$$

where $p(x)$ is is the unconditional likelihood like in (1).

The conditional information energy of Ω given by the fuzzy event A is as follows:

$$E(P(\Omega|A)) = \sum_{i=1}^{c}(P(\omega_i|A))^2, \tag{18}$$

where $P(\omega_i|A) = \frac{P(\omega_i)\int\limits_{\Re^d}\mu_A(x)p(x|\omega_i)dx}{P_m(A)}$.

The conditional information energy of Ω given the fuzzy information $\mathcal{A}$ is as follows:

$$E(\mathcal{A},\Omega) = \sum_{A\in\mathcal{A}} E(P(\Omega|A))P_m(A). \tag{19}$$

For such definition of conditional information energy the upper and lower bounds on probability of error, similarly as in [11], are given by:

$$\frac{1}{2}(1 - E(\mathcal{A},\Omega)) \leq Pe(\Psi_F^*) \leq (1 - E(\mathcal{A},\Omega)). \tag{20}$$

4.4 Numerical example

The aim of the experiment is to compare the estimation error of Bayesian classifiers calculated from (14) with the bounds of error classification obtained in terms of the information energy of fuzzy sets. Additionally, relationship between the estimation error (14) and information energy of fuzzy events (16) is introduced. These results are calculated for a full probabilistic information.

Let us consider the binary classifier with *a priori* probabilities $P(\omega_1) = P(\omega_2) = 0.5$. The class-conditional probability density functions are normal distributions in $\Re^1$ $p(x|\omega_1) = N(5.5, 1)$ and $p(x|\omega_2) = N(6.5, 1)$. In experiments, the following sets of fuzzy numbers were used:
case A

$$\mathcal{A} = \{A^1 = (-2,\, 0,\, 2)\,,\ A^2 = (0,\, 2,\, 4)\,, \ldots, A^8 = (14,\, 16,\, 18)\},$$

case B

$$\mathcal{A} = \{B^1 = (-1,\, 0,\, 1)\,,\ B^2 = (0,\, 1,\, 2)\,, \ldots, B^{16} = (14,\, 15,\, 16)\},$$

Tab. 1 shows the difference between the probability of misclassification for fuzzy and non fuzzy data in the Bayes optimal classification calculated from (14) and information energy of fuzzy information (16) for case A. The difference $(1 - E(\mathcal{A}, \Omega)) - Pe(\Psi^*)$ for this case is equal 0.126. The change of this difference in dependence from the parameter k is the poses the precision 0.0001. The Tab. 2 shows suitable results for the case B. The information energy of fuzzy information $W(\mathcal{B})$ is equal 0.2358 and the difference $(1 - E(\mathcal{B}, \Omega)) - Pe(\Psi^*)$ is equal 0.4105 with the precision 0.0001.

Table 1. The difference between the probability of misclassification $Pe(\Psi_F^*) - Pe(\Psi^*)$ and information energy of fuzzy information $W(\mathcal{A})$ for case A

		$p((x-k)\|\omega_1)$,		$p((x-k)\|\omega_2)$,		$k =$	
	0	0.5	1	1.5	2	2.5	3
$W(\mathcal{A})$	0.4093	0.4046	0.4000	0.4046	0.4093	0.4046	0.4000
$Pe(\Psi_F^*) - Pe(\Psi^*)$	0.0732	0.0390	0.0257	0.0390	0.0732	0.0390	0.0257

Table 2. The difference between the probability of misclassification $Pe(\Psi_F^*) - Pe(\Psi^*)$ for case B

		$p((x-k)\|\omega_1)$,		$p((x-k)\|\omega_2)$,		$k =$	
	0	0.25	0.5	0.75	1	1.25	1.5
$Pe(\Psi_F^*) - Pe(\Psi^*)$	0.0257	0.0120	0.0070	0.0120	0.0257	0.0120	0.0070

The parameter k shifts the discriminant functions $P(\omega_1)p((x-k)|\omega_1)$ and $P(\omega_2)p((x-k)|\omega_2)$. Fuzzy observations are represented by adequate fuzzy numbers. The following conclusions could be drawn from the experiment:

- the difference in the misclassification for fuzzy and crisp data does not depend only on the width fuzzy number,
- the position of the class-conditional pdf's in relation to the observed fuzzy features is the essential influence for the difference $Pe(\Psi_F^*) - Pe(\Psi^*)$,
- this difference is periodical, the period is equal a half of the width fuzzy number,
- the information energy of fuzzy events is periodical too, the period is equal a half of the width fuzzy number,
- the differences $Pe(\Psi_F^*) - Pe(\Psi^*)$ is exact, the difference based on information energy $(1-E(\mathcal{A},\Omega))-Pe(\Psi^*)$ is quite inaccurate estimation of the difference of error for fuzzy and crisp data.

5 Conclusion

In the present paper we have concentrated on the Bayes optimal classifier. Assuming a full probabilistic information we have presented the difference between the probability of misclassification for fuzzy and crisp data. Additionally, the received results are compared with the bound on the probability of error based on information energy.

Acknowledgements: This research is supported in part by The Polish State Committee for Scientific Research under the grant which is realizing in years 2010–2013.

References

1. Antos A, Devroye L, Gyorfi L (1999) Lower Bounds for Bayes Error Estimation. IEEE Trans. Pattern Analysis and Machine Intelligence, **21**, 643–645
2. Avi-Itzhak H, Diep T (1996) Arbitrarily Tight Upper and Lower Bounds on the Bayesian Probability of Error. IEEE Trans. Pattern Analysis and Machine Intelligence, **18**, 89–91
3. Burduk R, Kurzyński M (2006) Two-stage binary classifier with fuzzy-valued loss function. Pattern Analysis and Applications **9** No 4 353–358
4. Burduk R (2010) Classification error in Bayes multistage recognition task with fuzzy observations. Pattern Analysis and Applications **13** No 1 85–91
5. Hanas S (2001) On the Interval Approximation of a Fuzzy Number. Fuzzy Sets and Systems **122**, 353–356
6. Janickow C Z (1998) Fuzzy Decision Trees: Issues and Methods. IEEE Trans. Systems, Man, and Cybernetics B: Cybernetics **28**, 1–14
7. Kulkarni A (1978) On the Mean Accuracy of Hierarchical Classifiers. IEEE Transactions on Computers **27**, 771–776
8. Kuncheva L I (2004) Combining Pattern Classifier: Methods and Algorithms. John Wiley, New York

9. Kurzyński M (1988) On the Multistage Bayes Classifier. Pattern Recognition **21**, 355–365
10. Okuda T, Tanaka H, Asai K (1978) A formulation of fuzzy decision problems with fuzzy information using probability measures of fuzzy events. Information and Control **38**, 135–147
11. Pardo, J.A, Taneja I.J.: On the Probability of Error in Fuzzy discrimination Problems. Kybernetes **21** No. 6 (1992) 43–52
12. Pardo, L: Information Energy of a Fuzzy Event and a Fuzzy Events. IEEE Trans. on Systems, Man. and Cybernetisc **SMC-15** No. 1 (1985) 139–144
13. Pardo, L, Menendez M.L.: Some Bounds on Probability of Error in Fuzzy Discrimination Problems. European Journal of Operational Research **53** (1991) 362–370
14. Pedrycz W (1990) Fuzzy Sets in Pattern Recognition: Methodology and Methods. Pattern Recognition **23**, 121–146
15. Stańczyk U.: Dominance-Based Rough Set Approach Employed in Search of Authorial Invariants, Advances in Intelligent and Soft Computing **57**, Springer-Verlag, Berlin Heidelberg (2009) 293–301
16. Supriya K.D., Ranjit B., Akhil R.R.: An application of intuitionistic fuzzy sets in medical diagnosis. Fuzzy Sets and Systems 117 No 2 (2001) 209–213
17. Woźniak, M.: Experiments on linear combiners. Advances in Soft Computing **47** Springer Berlin (2008) 445–452
18. Zadeh L A (1968) Probability Measures of Fuzzy Events. Journal of Mathematical Analysis and Applications **23**, 421–427