Combined Approach to Pattern Classification in Parametric Case

MICHAL WOZNIAK
Wroclaw University of Technology
Chair of Systems and Computer Networks
Wyb. Wyspianskiego 27, 50-370 Wroclaw
POLAND

Abstract: This paper is devoted to the methods for combining heterogeneous sets of learning data: set of training examples and set of IF-THEN rules with unprecisely formulated weights. Adopting the probabilistic (Bayes) model of recognition task and assuming known form of class conditional probability density functions (CPDFs) with unknown parameters, the recognition algorithm via fusion of both sets of data is presented. Proposed concept of combining of input data consists in treating of both sets as sources of information about unknown parameters of CPDFs, which leads to the modified maximum likelihood (ML) method of parametr estimation. In proposed procedure the likelihood function is maximized taking into account constraints provided by the set of rules. A series of numerical examples with computer generated data for several cases which differ in form and number of rules is considered. To find feasible solution of ML problem two approaches were employed. The first method uses the Kuhn-Tucker conditions for the nonlinear problem with inequality constraints, the second one however is approximated procedure which does not guarantee the optimality of result. For each solution the estimation error, i.e. distance between values of estimator and parameter is calculated as a measure of its quality, which allows us to rank procedures and to imply some practical conclusions.

Key–Words: Pattern recognition, Parametric case, Maximum likelihood method

1 Introduction

During the two past decades the fusion of various sources of knowledge was firmly established as a practical and effective solution for difficult pattern recognition tasks ([1], [2], [3]). This idea is established using classifier combination approach, which in the literature is known under many names: hybrid methods, decision combinations, classifier fusion, mixture of experts, modular systems, to name only a few ([4]).

Fusion of the classifiers outputs is the very promised way of constructing combined algorithms. Most of the research on classifier ensambles is concerned with generating ensambles by using a single learning model. Different classifiers are received by manipulating the training set, or the input features, and next their decisions are combined in some way (typically by voting) to classify new patterns. Another approach is to generate classifiers by applying different learning algorithms to a single data set ([4]).

For the probabilistic model of recognition task and Bayesian decision theory, fusion of classifiers denotes combining of estimators of *posterior* probabilities of classes, which are produced by simple classifiers on the base of their input data. There are many ways known in literature of fusion of *posterior* probability estimators. As an example, one has to mention the works based on class-conscious combiners [4] where the discriminant functions of the combined classifier are obtained as the average values of *posterior* probabilities of simple classifiers. In another methods discriminant function of combined classifier is obtained via weighted sum of the estimators of *posterior* probabilities of simple classifiers ([5], [6], [7], [8]) and the concrete procedures differ with concepts of weight coefficients.

In the present paper the novel classification problem with probabilistic model is discussed, in which we assume known form of class conditional probability density functions (CPDFs) with unknown parameters and as an input data the set of training examples (learning set) and the set of expert rules are considered. The main question is how to utilize the information contained in the both sets to obtain good estimates for the unknown parameters. To solve this problem we propose to apply the modified maximum likelihood method of parameter estimation, in which the likelihood function is maximized taking into account constraints provided by the set of rules. This approach, i.e. the use of additional source of information contained in rules leads to the better estima-

tion results in comparison with maximum likelihood method without constraints.

This paper is a sequel to the authors earlier publications ([8], [9], [10]) and it yields an essential extension of the results included therein and dealing with combining rule-based and sample-based classifiers.

The contents of the paper are as follows. In section 2 we introduce necessary background and formulate the parametric case of classification problem with probabilistic model and for input data contained in learning set and expert rules. In section 3 combined classifier for the problem in question is proposed, which construction is reduced to the maximization of likelihood function with constraints provided by expert rules. Section 4 describes numerical examples and yields results of comparative analysis of two optimization procedures for different form and number of expert rules. Finally, conclusions are presented in section 5.

2 Preliminaries and the Problem Statement

Among different approaches to the uncertainty management in computer-aided recognition systems, the statistical decision theory is still an attractive and effective method ([12], [13], [14]). This theory assumes that both the vector of features describing recognized pattern $x \in \mathcal{X} \subseteq \mathcal{R}^d$ and its class number $j \in \mathcal{M} = \{1, 2, ..., M\}$ are observed values of a pair of random variables $(\mathbf{X}, \mathbf{J})$, respectively. Its probability distribution is given by *prior* probabilities of classes $p_j = P(\mathbf{J} = j)$ and class-conditional probability density functions (CPDFs) of $\mathbf{X}$ - $f_j(x) = f(x \mid j)$ ($x \in \mathcal{X}$, $j \in \mathcal{M}$).

In pattern recognition a function $\psi(x) : \mathcal{X} \to \mathcal{M}$ is called a classifier. If we know the *priors* and CPDFs then we can design the optimal (Bayes) classifier ψ^*, minimizing the probability of misclassification, which makes decision according to the following rule:

$$\psi^*(x) = i \text{ if } p_i(x) = \max_{k \in \mathcal{M}} p_k(x), \tag{1}$$

where *posterior* probabilities $p_j(x) = P(\mathbf{J} = j \mid x)$ can be calculated from the Bayes formula, viz.

$$p_j(x) = \frac{p_j \, f_j(x)}{f(x)}. \tag{2}$$

Although in practical pattern recognition problems we rarely have complete knowledge about *priors* and CPDFs, sometimes however, an advance vague information about the probability distribution of $(\mathbf{X}, \mathbf{J})$ is available. This information can have various nature and can be quite specific (e.g. „in the first class features are normal“, or „probability of the second class is thought to be a monotone function of $x \in \mathcal{R}$“, or „for $x < 0$ probability of the third class is almost constant“), but very often it refers to forms of CPDFs, i.e. describes the class of functions (or probability distributions) which CPDFs belong to.

Let us now consider exactly such a case, i.e. we assume that for each class $j \in \mathcal{M}$, CPDF $f_j(x)$ has known parametric form and is uniquely determined by the values of the parameter vector θ_j. To emphasise this fact we will further write $f_j(x)$ as $f_j(x; \theta_j)$. Parameters θ_j are constant but unknown and *prior* probabilities p_j are assumed to be unknown as well.

Instead, suppose we have two qualitatively different kinds of data which contain hidden information on unknown prameters of probability distribution of $\mathbf{J}$ and $\mathbf{X}$. Let us present form of available information we may use.

1. Learning set:

$$S = \{(x_1, j_1), (x_2, j_2), ..., (x_N, j_N)\}, \tag{3}$$

where x_i denotes the feature vector of the i-th learning pattern and j_i is its correct classification.

Additionally, let $S_i = \{x_{i1}, x_{i2}, ..., x_{iN_i}\}$ denotes the set of learning patterns from the i-th class. Samples in S_i are assumed to be drawn independently from a distribution with parameters θ_i and they do not yield information about θ_j, $j \neq i$.

2. Expert rules:

$$R = \{R_1, R_2, ..., R_M\}, \tag{4}$$

where

$$R_i = \{r_i^{(1)}, r_i^{(2)}, ..., r_i^{(K_i)}\}, \;\; i \in \mathcal{M}, \;\; \sum K_i = K \tag{5}$$

denotes the set of rules connected with the i-th class. The rule $r_i^{(k)}$ has the following general form:

IF $x \in D_i^{(k)}$ **THEN** $\mathbf{J} = i$ **WITH** *posterior* probability greater than $\underline{p}_i^{(k)}$ and less than $\overline{p}_i^{(k)}$.

Equivalently, the rule $r_i^{(k)}$, treated as a source of information on probability distribution of $\mathbf{J}$ and $\mathbf{X}$, determines the following inequalities for x belonging to the rule-defined region $D_i^{(k)} \subset \mathcal{X}$:

$$\underline{p}_i^{(k)} \leq p_i(x) \leq \overline{p}_i^{(k)}. \tag{6}$$

Now our purpose is to construct the recognition algorithm

$$\psi_{SR}(x) = i, \tag{7}$$

which uses information provided by the learning set S and the set of expert rules R to classify a pattern on the basis of its features x.

3 Combined Classifier

For the problem in question, one obvious and conceptually simple method is to construct estimates $\hat{p}_j$ and $\hat{\theta}_j$ of unknown parameters p_j and θ_j ($j \in \mathcal{M}$) from available data and next to use them in the optimal Bayes algorithm (1) as though they were correct. In other words, we consult an expert (or experts) and have access to a database of examples observed in the past, to try to reconstruct Bayes classifier ψ^*. This idea leads to the following combined classifier:

$$\psi_{SR}(x) = i \text{ if } \hat{p}_i f_i(x;\hat{\theta}_i) = \max_{k\in\mathcal{M}} \hat{p}_k f_k(x;\hat{\theta}_k). \quad (8)$$

The task of parameter estimation is a classical one in statistics and it can be approached in several methods. In problem at hand, the estimation of the *prior* probabilities yields no serious difficulties and usually is performed according to the following formula:

$$\hat{p}_i = \frac{N_i}{N}, \ i \in \mathcal{M}. \quad (9)$$

The main question is how to utilize the information contained in the sets S and R to obtain good estimates for the unknown parameters $\theta_1, \theta_2, ..., \theta_M$ associated with each class. To solve this problem we propose to apply on the base of set S the maximum likelihood (ML) method which additionally respects constraints provided by the set R.

Applying this approach for ith class ($i \in \mathcal{M}$), one must first determine the joint CPDF of data set S_i, which expressed as a function of unknown parameters θ_i is called the likelihood function. Under adopted assumptions (see decsription of learning set in the previous section) the likelihood function can be determined as follows ([16]):

$$L_i(\theta_i) = f(S_i;\theta_i) = \prod_{n=1}^{N_i} f_i(x_{in};\theta_i) \quad (10)$$

Since the ML estimate is „the most likely" value given the observed data, in the next step of ML procedure we maximize the likelihood function with respect to parameter vector θ_i. Feasible domain of solutions is determined by the set of inequalities provided by set R, which from (2), (6) and (9) can be explicitly expressed as follows:

$$\underline{p}_i^{(k)} \le \frac{\frac{N_i}{N} f_i(x;\theta_i)}{\sum_{n=1}^{M} \frac{N_n}{N} f_n(x;\theta_n)} \le \overline{p}_i^{(k)}, \quad (11)$$

$$x \in D_i^{(k)}, \ k = 1, 2, ..., K_i.$$

Thus, the problem of construction classifier (7), i.e. combined learning procedure on the base of learning set and expert rules is equivalent to the solution of the constrained global (nonlinear in general case) optimization problem of the form:

$$\text{Find } \hat{\theta}_i \text{ such that } L_i(\hat{\theta}_i) =$$

$$= \max_{\theta_i} L_i(\theta_i) \text{ subject to (11), } i \in \mathcal{M}. \quad (12)$$

Unfortunately let us note that, though optimization problems are formulated separately for each class (for each parameter vector θ_i), constraints involve simultaneously all unknown parameters $\theta_1, \theta_2, ..., \theta_M$, which causes optimization problems (12) mutually dependent. This fact justifies considertion of optimization problem with common objective function, which according to the concept of *weighting method* [15], we adopt as the sum of likelihood functions for particular tasks, i.e. $L(\theta) = \sum_{i\in\mathcal{M}} L_i(\theta_i)$, where $\theta = (\theta_1, \theta_2 ... \theta_M)$. Thus, we shall reformulate (12) as follows:

$$\text{Find } \hat{\theta} \text{ such that } L(\hat{\theta}) = \max_{\theta} L(\theta) \text{ subject to (11).} \quad (13)$$

To find solution of (13) we can use the Kuhn-Tucker conditions for the nonlinear problem with inequality constraints [15]. Unfortunately, in general case, the set of feasible solutions determined by the inequalities (11) does not need be convex set. This observation provides serious difficulties in obtaining solution, even in quite simple cases.

In next section we present several examples which illustrate various cases of expert rules leading to the different forms of sets of feasible solutions. Obtained results make possible to assess usefulness of additional knowledge in the form of expert rules in the problem of estimation of unknown parameters of probability distribution.

4 Numerical Examples

Let us consider two-class pattern recognition task. Scalar feature has normal distribution in both classes, i.e.:

$$f_1(x) \sim N(m_1, \sigma_1), \ f_2(x) \sim N(m_2, \sigma_2). \quad (14)$$

We assume $\sigma_1 = \sigma_2 = 1$ and mean values are unknown parameters, i.e. $\theta_1 = m_1$ and $\theta_2 = m_2$. In order to generate learning set (3) we assumed $N = 15$, $m_1 = 0, m_2 = 1$ and $p_1 = 2/3$. Using random numbers generator in Maple 10 environment we received the following learning patterns from the first and the second class, respectively: $S_1 = \{0.853,$ -0.591, 0.578, 0.226, 0.723, -0.809, -0.516, 0.098, 0.514, $1.064\}$ and $S_2 = \{0.185,$ 0.203, 1.117, 2.109, $0.374\}$.

First, we use learning set as a random sample to the estimation of unknown mean values. The ML method without constraints leads to the following results:

$$\hat{m}_1 = 0.214, \quad \hat{m}_2 = 0.797. \tag{15}$$

The course of likelihood functions is depicted in Fig.1. Let us note that L_1 and L_2 are unimodal functions - this property will be utilized in further examples.

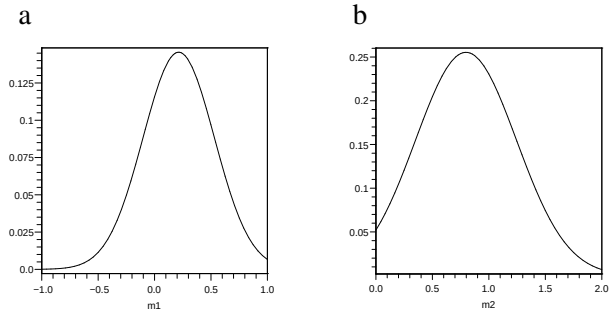

Figure 1: The course of likelihood function $L_1(m_1)$ (a) and $L_2(m_2)$ (b)

Now we discuss the case when both learning set and set of expert rules are available. Let us start from the simplest case in which the set R contains one rule only with rule-defined region equal to one-point set, i.e. $D = x_0$ Without any loss of generality, we suppose that rule is connected with the first class. Thus now, inequalities (6) reduce to the form

$$\underline{p} \leq p_1(x_0) \leq \overline{p}. \tag{16}$$

Hence and from (14), after simple calculations we get the following constraints (11):

$$2ln(\frac{\underline{p}}{1-\underline{p}} \cdot \frac{1-p_1}{p_1}) \leq (x_0 - m_2)^2 - (x_0 - m_1)^2 \leq$$

$$\leq 2ln(\frac{\overline{p}}{1-\overline{p}} \cdot \frac{1-p_1}{p_1}). \tag{17}$$

The above two inequalities determine in the space $\mathcal{R}^2$ set of feasible solutions of (13), which shape and size strictly depends on values of bounds $\overline{p}, \underline{p}$ and accuracy of determining *posterior* probability in rule $\Delta = \overline{p} - \underline{p}$. Some examples for values $\underline{p} = 0.5, \overline{p} = 0.9, \Delta = 0.4$ (solid line), $\underline{p} = 0.67, \overline{p} = 0.85, \Delta = 0.18$ (dash line) and $\underline{p} = 0.7, \overline{p} = 0.8, \Delta = 0.1$ (dot line) are depicted in Fig.2a. In next examples we restrict the space of feasible solutions to the case of positive values of m_2.

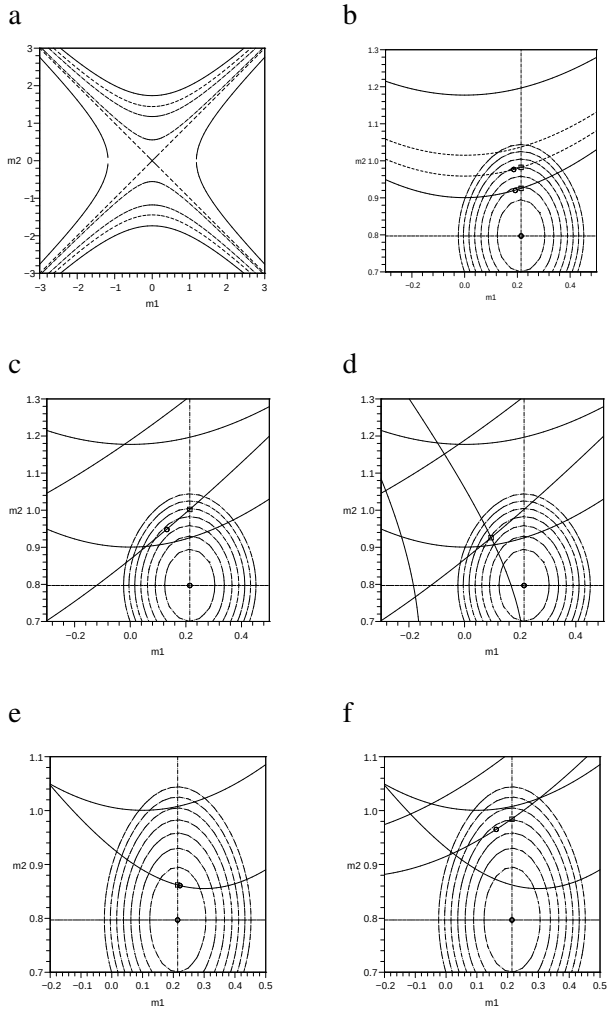

Figure 2: Illustration of example (details in the text)

Fig.2b. presents the sets of feasible solutions determined by one point-rule with $x_0 = 0$ for two cases of bounds and accuracy: $\underline{p} = 0.75, \overline{p} = 0.8, \Delta = 0.05$ (solid line) and $\underline{p} = 0.76, \overline{p} = 0.78, \Delta = 0.02$ (dash line). The objective function is visible in the form of elipsoid contourlines.

As alternative we apply approximate method of solution search, which does not guarantee the optimal solution. The idea is very simple and does not yield any conceptual and technical problems. In this method, if result of ML method (without constraints)

Table 1: Results of estimation of mean values for different cases and methods (Abbreviations: WCM - ML method without constraints, 1PRw(n) - one point-rule (wide)(narrow), 2(3)PR - two (three) one-point rules, 1(2)IR - one (two) interval rule(s))

Method	Optimal Solution			Suboptimal Solution		
	$\hat{m}_1$	$\hat{m}_2$	Q	$\hat{m}_1$	$\hat{m}_2$	Q
WCM	0.214	0.797	0.417	-	-	-
1PRw	0.192	0.920	0.272	0.214	0.926	0.288
1PRn	0.186	0.977	0.209	0.214	0.982	0.232
2PR	0.131	0.948	0.183	0.214	1.002	0.216
3PR	0.095	0.926	0.169	-	-	-
1IR	0.222	0.861	0.361	0.214	0.862	0.352
2IR	0.163	0.965	0.198	0.214	0.984	0.230

is situated outside of the set of feasible solutions, one value of parametr estimator (say $\hat{m}_1$) remains unchanged, however we vary the second one so as to reach the nearest constraint (see properties of likelihood functions).

In the next two cases, to point-rule in $x_0 = 0$ with $\underline{p} = 0.75, \overline{p} = 0.8, \Delta = 0.05$ we add the second one in $x_0 = -1.3$ with $\underline{p} = 0.9, \overline{p} = 0.95, \Delta = 0.05$ and the third rule in the point $x_0 = 0.6$ with $\underline{p} = 0.6, \overline{p} = 0.65, \Delta = 0.05$. The sets of feasible solution determined by these two and three point-rules, their locations relative to objective function and solutions given by the optimal and suboptimal methods are presented in Fig.2c and Fig.2d, respectively.

In the further examples we consider the case of expert rules with rule-defined region equal to interval on $\mathcal{R}$, i.e. $D = [a, b]$. Since in the example at hand, *posterior* probability of the first class is decreasing function, it is obvious that constraints (6) reduce to the two following inequalities:

$$\underline{p} \leq p_1(b),\ p_1(a) \leq \overline{p}, \tag{18}$$

which explicitly determine the set of feasible solutions of (13).

Results for one rule ($D = [0.1, 0.3], \underline{p} = 0.7, \overline{p} = 0.75, \Delta = 0.05$) and for the case with added the second rule ($D = [-0.5, -0.3], \underline{p} = 0.8, \overline{p} = 0.85, \Delta = 0.05$) are depicted in the Fig.2e and Fig 2f, respectively.

Estimators of mean values for all examples are presented in Table 1. For each pair of estimators $(\hat{m}_1, \hat{m}_2)$ additionally an error Q was calculated:

$$Q = \mid m_1 - \hat{m}_1 \mid + \mid m_2 - \hat{m}_2 \mid, \tag{19}$$

which evaluates the quality of estimate and allows to compare obtained results for different cases and estimation procedures.

These results and values of criterion Q imply the following conclusions:

1. The ML method without constraints (WCM) is worse than those that used additional information contained in expert rules, even for suboptimal approach. This confirms the effectiveness and usefulness of the conceptions and procedure construction principles presented above for the needs of parametric pattern recognition.

2. Results of method with one point-rule strictly depend on the accuracy Δ of determining *posterior* probability in rule. Less value of Δ causes less value of Q, i.e. we obtain the better result.

3. There occurs a common effect within each kind of rules group (point-rule and interval-rule) and for both methods: bigger number of rules causes the better estimation.

4. Although the optimal algorithm yields always better results than suboptimal one, in many cases there are no essential difference between both methods. This fact allows to consider suboptimal approach as an interesting alternative in the optimization problem in question. We must remember however, that suboptimal approach not always leads to the constructive results (see e.g. case 3PR).

5 Conclusion

This paper presents probabilistic approach to the combinig of learning set and the set of IF-THEN rules for the parametric case. In proposed concept of input data fusion, both sets are treated as sources of information on unknown parameters, which - in a natural way - leads to the modified ML method of parametr estimation. In proposed procedure, on the base of learning set the likelihood function is formulated, and next it is maximized taking into account constraints provided by the set of rules.

We have considered a series of numerical examples with computer generated data for several cases which differ in form and number of rules. For each example two optimization procedures were employed to obtain estimators of unknown parameters. Furthermore, for each solution the estimation error, i.e. distance between values of estimator and parametr was calculated. The ranking of procedures and cases cannot be treated as one having the ultimate character because the scope of numerical examples warns us

against its uncritical use. However, although the outcome may be different for other tasks, the presented examples demonstrate the usefulness of information contained in expert rules and may suggest some perspectives for practical applications.

References:

[1] Xu L., Krzyzak A., Suen Ch.Y.: Methods of Combining Multiple Classifiers and Their Applications to Handwriting Recognition. IEEE Trans. on SMC, Vol.22. (1992) 418-435

[2] Kittler J., Duin R., Matas J. (1998) On combining classifiers, IEEE Trans. on PAMI, vol. 20 : 226-239

[3] Hansen L.K., Salamon P.: Neural Networks Ensembles. IEEE Trans. on PAMI, Vol. 12. (1990) 993-1001

[4] Kuncheva L.I.: Combining pattern classifiers: Methods and algorithms. Wiley-Interscience, New Jersey (2004)

[5] Ting K.M., Witten I.H.: Issues in stacked generalization, Journal of Artificial Inteligence Research, Vol. 10 (1999) 271-289

[6] Hashem S.: Optimal linear combinations of neural networks, Neural Networks, Vol. 10 (1997) 599-614

[7] Yager R.R.: On Order Weighted Averaging Operators in Multicriteria Decision Making, IEEE Trans. On Systems, Man, and Cybernetics, Vol. 18 (1988) 183-193

[8] Kurzynski M., Sas J., Blinowska A.: Rule-Based Medical Decision-Making with Learning, Proc. 12th World IFAC Congress, Vol. 4, Sydney (1993) 319-322

[9] Kurzynski M., Wozniak M.: Rule-Based Algorithms with Learning for Sequential Recognition Problem, Proc. 3rd Int. Conf. Fusion 2000, Paris (2000) 10-13

[10] Kurzynski M.: Combining rule-based and sample-based classifiers - probabilistic approach, Proc. Int. Conf. Brain, Vision and Artificial Intelligence, LNCS Vol.3704 (2005) 298-307

[11] Chow C.K.: Statistical independence and threshold functions. IEEE Trans. on Electronic Computers, Vol.16. (1965) 66-68

[12] Jain A.K., Duin P.W., Mao J.: Statistical Pattern Recognition: A Review. IEEE Trans. on PAMI, Vol. 22. (2000) 4-37

[13] Devroye L., Gyorfi P., Lugossi G.: A probabilistic theory of pattern recognition. Springer Verlag, New York (1996)

[14] Duda R., Hart P., Stork D.: Pattern classification. Wiley-Interscience, New York (2001)

[15] Miettinen K.: Multiobjective optimization. Kluwer Academic Publisher, Boston (1999)

[16] Sachs L.: Applied statistics. A handbook of techniques. Springer-Verlag, New York (1984)