Synthetic pattern recognition methods based on the representational minimum description length principle

A.S. Potapov

Saint Petersburg State University of Information Technologies, Mechanics and Optics
49 Kronverksky ave., Saint Petersburg, 197101, Russia
E-mail: pas.aicv@gmail.com

ABSTRACT

The minimum description length principle (MDL) is considered in context of pattern recognition problem. It is shown that existent applications of the MDL principle use heuristic coding schemes in order to calculate code length in contrast to the theoretic MDL principle based on the notion of algorithmic complexity. The representational MDL principle is proposed as an extension of the MDL principle that replaces heuristic schemes with the notion of information representation with strict measure of quality. Pattern recognition problem is reduced to the task of choosing such the optimal representation that minimizes summed description length of patterns in a training set. As a result, synthetic methods with automatic selection between particular methods are proposed.

Keywords: pattern recognition, representational minimum description length, model selection

1. INTRODUCTION

Pattern recognition is a well-studied, but still not entirely solved problem. One of the most difficult questions here is the question of selection between recognition models of different complexity, e.g. between nonlinear discrimination functions with different number of coefficients or between Gaussian mixtures with different number of components. The minimum description length (MDL) principle solves this problem to some extent[1,2].

This principle deals with the problem of induction (i.e. model selection on the base of given data) and states that, among the given set of models, one should choose the model that minimizes the sum of the length (in bits) of the data encoded with the help of this model and of the length (in bits) of the model description[3]. In theoretic studies[4], description lengths are calculated on the base of the algorithmic complexity (length of the minimum universal Turing machine program, which generates given data). That is, the MDL principle determines well-grounded tradeoff between the likelihood (or precision) of the model to the training sample and the complexity of the model itself.

Different pattern recognition methods can be improved by using the MDL principle[1,2,5]. However, these methods use heuristic (ungrounded) coding schemes in order to calculate the description length criterion. For example, arbitrary basis functions can be used to represent discrimination functions, or an arbitrary distance function can be used as the similarity measure in the template or prototype matching method. This creates the gap between theoretically grounded MDL principle and its practical applications in pattern recognition. Thus, another problem arises: how to prove the use of a certain coding scheme or how to choose the best one?

In order to solve this problem one should introduce a meta-level in the task of induction, i.e. the task of selection between different sets of models. We use a notion of *representation* instead of set of models. A certain representation is a symbolic system, within which descriptions of models are realized. In this paper formal definition of this notion is proposed and the MDL principle is extended with its use. This helps to fill the gap between theoretical and practical studies of the MDL principle and to derive the criterion for choosing between different model spaces (coding schemes or representations). As a result, synthetic methods with automatic selection between particular methods are proposed. Increase of recognition rate of synthetic methods is shown.

2. PATTERN RECOGNITION PROBLEM

The following components are considered in the task of pattern recognition:

- Feature (or description) space X; each object supposed for recognition is represented as a point in this space $\mathbf{x} \in \mathrm{X}$;
- Set of classes $\mathrm{A} = \{a_1, ..., a_d\}$ consisting finite number of elements;
- mapping $\varphi_0 : \mathrm{X} \to \mathrm{A}$ (true membership function) that splits the feature space into n regions; mapping φ_0 is unknown in the task of recognition;
- initial data (training set); it can contain whether the set $\{(\mathbf{x}_i, \mu_i)\}_{i=1}^n$, where $\mathbf{x}_i \in \mathrm{X}$, $\mu_i = \varphi_0(\mathbf{x}_i)$, or the set $\{\mathbf{x}_i\}_{i=1}^n$ depending on task (supervised and unsupervised learning).

The task is to find such the mapping $\varphi : \mathrm{X} \to \mathrm{A}$ that the norm $\|\varphi - \varphi_0\|$ is minimal, that is the mapping φ should precisely predict values of φ_0 on the set X. We consider the recognition problem, in which $\mathrm{X} = R^N$ (N is the number of features).

Criterion $\|\varphi - \varphi_0\|$ is incorrect, because true function φ_0 is unknown, so one should also approximate this criterion on the base of initial data. It should be pointed out that calculation of $\|\varphi - \varphi_0\|$ only on points of learning set, which seems to be very natural, leads to the problem of overlearning: mapping φ can precisely approximate true function φ_0 at points of training set, but at the same time errors can be very large at the other points. Discriminative power of mapping φ chosen on the base of this criterion can be very low. The problem of construction of an adequate criterion can be solved on the base of the MDL principle.

2. REPRESENTATIONAL MINIMUM DESCRIPTION LENGTH PRINCIPLE

2.1. The minimum description length principle

The MDL principle can be formally introduced in the following way. Let U be the universal Turing machine (UTM). Let prefix algorithmic complexity of the binary string β be $K_U(\beta) = \min_{\alpha}[l(\alpha) \mid U(\alpha) = \beta]$, where $l(\alpha)$ is the length of the program α. Index U will be omitted for simplification of notation, when it is clear from context. Program for UMT can be considered as a model of the source, which generated data β.

String α can be represented as concatenation of two strings $\alpha = \mu\delta$, where μ is interpreted as the program itself (the model or regular component), and δ is the initial data for this program (random component). Then it can be written

$$K(\beta) = \min_{\mu}\left[l(\mu) + \min_{\delta}\left[l(\delta) \middle| U(\mu\delta) = \beta \right] \right] = \min_{\mu}\left[l(\mu) + K(\beta \mid \mu) \right], \quad (1)$$

where $K(\beta \mid \mu)$ is the conditional complexity of the string β with the given string μ.

Model selection in the framework of this approach is performed by inverting the problem of optimal coding[6] that is performed on the base of knowledge of the data source. Here, the best model is determined as the model that provides optimal coding (1):

$$\mu^* = \arg\min_{\mu}\left[K(\beta \mid \mu) + l(\mu) \right]. \quad (2)$$

Equation (2) corresponds to the minimum description length principle: the best model μ to describe data β is the model that provides the minimum value of the sum

- the length of the model $l(\mu)$;
- the length of the data described with the help of the model $K(\beta \mid \mu)$.

Application of the equation (2) to the pattern recognition problem is not straightforward. String β stands for entire training set, and model μ describes arbitrary regularities in the data, but it is not a model for recognizing new patterns. This is the reason, why existent MDL-based recognition methods[1,2,5] don't use the equation (2), but instead rely on general verbal definition of the MDL principle and heuristic coding schemes to calculate the length of the model and the length of the data in the context of the pattern recognition. We propose the representational MDL (RMDL) principle to deal with this problem more formally.

2.2. The representational minimum description length principle

Consider the problem of unsupervised learning with the given training set $\{\mathbf{x}_i\}_{i=1}^n$. The main property of the recognition model is that it is applied to each pattern independently. Consider the length of independent descriptions of the patterns. In this case, one should solve *M* problems $\mu_i^* = \arg\min_{\mu}\left[K(\mathbf{x}_i \mid \mu) + l(\mu) \right]$. The summed description length is expressed as

$\sum_{i=1}^{n} K(\mathbf{x}_i) \geq K(\mathbf{x}_1\mathbf{x}_2...\mathbf{x}_n)$. Inequality will be strict, if pattern contain any mutual information.

Moreover, length of independent descriptions greatly depends on the choice of UTM. Consider two UTM's: *U* and *V*. It is known[4] that for any two UTM's such the string ν exists that $(\forall\alpha) U(\nu\alpha) = V(\alpha)$. Consequently, $K_U(\alpha) \leq K_V(\alpha) + l(\nu)$, *i.e.*

$K_U(\mathbf{x}_1\mathbf{x}_2...\mathbf{x}_n) \leq K_V(\mathbf{x}_1\mathbf{x}_2...\mathbf{x}_n) + l(\nu)$.

Value of $l(\nu)$ is constant, so relative difference in algorithmic complexities for different UTM's reduces with increase of training set. This is the reason to ignore the choice of UTM. However, this is not true for the case of independent description of pattern:

$\sum_{i=1}^{n} K_U(\mathbf{x}_i) \leq \sum_{i=1}^{n} K_V(\mathbf{x}_i) + nl(\nu)$.

We introduce the following notion of representation in order to overcome these difficulties.

Definition. Let *representation* for the set of patterns X be defines as such the program *S* for the UTM *U* that the condition $(\forall \mathbf{x} \in \mathrm{X})\left(\exists \mu, \delta \in \{0,1\}^*\right) U(S\mu\delta) = \mathbf{x}$ is true. String $\mu\delta$ will be referred as *description* of the pattern $\mathbf{x}$ within the representation *S*.

Thus, two problems can be set: construction of description of a pattern within given representation and construction of representation on the base of training set. We propose *representational MDL* (RMDL) principle, which consists of two parts corresponding to these two problems.

1. The best model μ of the data element β (which can be patterns $\mathbf{x}_i$ or pairs $(\mathbf{x}_i, \mu_i)$ depending on the task) within the given representation S is the model, for which the following sum is minimized:
   - the length of the model $l(\mu)$;
   - the length of the data element described within the representation and the model $K(\beta \mid S\mu) = \min_{\delta}\left(l(\delta) \mid U(S\mu\delta) = \beta\right)$. Let $K(\beta \mid S\mu)$ be denoted as $K_S(\beta \mid \mu)$.

   The model selection is performed on the base of the following equations

$$L_S(\mathbf{x}, \mu) = K_S(\mathbf{x} \mid \mu) + l(\mu) \text{ and } \mu^* = \arg\min_{\mu} L_S(\beta, \mu). \quad (3)$$

2. The best representation S for the given set of patterns $D = \{\beta_i\}_{i=1}^n$ is the representation, for which the following sum is minimized:
   - the length of the representation $l(S)$;
   - sum of lengths of the best descriptions of patterns within the representation $\sum_{i=1}^{n} K_S(\beta_i) = \sum_{i=1}^{n} \min_{\mu} L_S(\beta_i \mid \mu)$.

   The representation selection is performed on the base of the following equations

$$L(D, S) = l(S) + \sum_{i=1}^{n} K_S(\beta_i) \text{ and } S^* = \arg\min_{S} L(D, S). \quad (4)$$

The MDL principle places the emphasis on the way of specifying the criterion of the model selection. However, it can be seen from the equation (3) that the most influence on the model selection is exerted by the used representation. Thus, if the best representation is selected from the severely bounded set containing no adequate solution, data description within this representation will not be effective in spite of the correct criterion of model selection. Analysis of representations becomes the main question to be investigated in order to extend capabilities of pattern recognition methods.

3. SYNTHETIC PATTERN RECOGNITION METHODS

3.1. Pattern recognition as selection of the best representation

The RMDL principle can be formulated for any mass problem differing by details of representations. Structure of representations for the pattern recognition tasks can be specified in the following way. Let $S_1, \ldots, S_d$ be representations, each of which should be determined for a single class. Then recognition of a single object $\mathbf{x}$ is the process of selection of a class with the representation yielding the best description of the object:

$\mu^* = \arg\min_{\mu=1,\ldots,d} L_S(\mathbf{x}, \mu) = \arg\min_{\mu=1,\ldots,d} K_{S_\mu}(\mathbf{x})$, where μ is the index of a class.

The problem of construction of a decision rule on the best of training sample is reduced to the task of construction of d representations. There can be two tasks depending on the structure of the training sample.

1. Both patterns $\mathbf{x}_i$ and corresponding classes $a_i = \varphi_0(\mathbf{x}_i)$ are given in the task of supervised learning. One should modify the RMDL criterion (4) in this case. It can be done in two different ways. The first way (which will be referred to as *inductive recognition*) is

to search for the best simultaneous description of patterns and corresponding indices class by minimizing the value $L(\{\mathbf{x}_i,\mu_i\}_{i=1}^n,S)$. The second way (which will be referred to as *predictive recognition*) is to describe only indices of classes assuming that patterns are given a priori, i.e. to minimize the value $L(\{\mu_i\}_{i=1}^n,S\,|\,\{\mathbf{x}_i\}_{i=1}^n)$.

2. Patterns $\mathbf{x}_i$ without classes are given in unsupervised learning. The task is to choose the best representation consisting of *d* partial representations, where *d* can also be unknown.

The full set of algorithms is not used in any practical pattern recognition method. Instead, the solution is chosen from a restricted parametric family of representations. This is the reason for these methods to be computationally effective. However, restrictions laid on the set of representations are usually very strong and not strictly grounded. As the result, each method of pattern recognition applicable only in certain cases depending on the characteristics of patterns distribution in the feature space. Since applicability of methods is usually not known a priori, one can propose an idea to select automatically between solutions obtained within several different methods. In the context of our approach, it just means that the involved set of representations is extended towards the algorithmically full set by combining several particular parametric families. Such the automatic selection can be performed on the base of the RMDL principle.

3.2. Synthetic representations

Let a parametric family of representations be denoted as $S\mathbf{w}$, where *S* is the program for UTM that accepts as its input the parameter vector $\mathbf{w}$ and description $\mu\delta$ of a pattern $\mathbf{x}$: $U(S\mathbf{w}\mu\delta)=\mathbf{x}$. The program *S* is constant within a certain method of pattern recognition. Parameters $\mathbf{w}$ are selected on the base of the learning sample, and the model μ is selected for each pattern. Any pattern recognition method can be described in this form. Applying the RMDL principle one can obtain

$$L(D,S)=l(S)+l(\mathbf{w})+\sum_{i=1}^{n}K_{S\mathbf{w}}(\beta_i)\,. \tag{5}$$

If the program *S* is constant, the term $l(S)$ can be ignored, and we can derive some existent applications of the classic MDL to the pattern recognition (see, for example, papers[1,2,5]). However, heuristically introduced coding schemes were used in these applications in order to compute the description length, because the important component *S* was not specified in explicit form. Probably, it's unavoidable to use such the coding schemes, because the search for the solution in the algorithmically full set is undecidable problem. However, as it was mentioned above, one can select not only the best parameters $\mathbf{w}$ for the given family $S\mathbf{w}$, but also to select the best family *S* from certain restricted set.

Definition. Let $S^{(1)}\mathbf{w}^{(1)},...,S^{(t)}\mathbf{w}^{(t)}$ be the set of parametric families of representations. Each representation corresponds to a certain pattern recognition method. *Synthetic pattern recognition method* is the method, in which optimal value of parameters $\mathbf{w}^{(i)}$ is determined for each family $S^{(i)}\mathbf{w}^{(i)}$, and the best family is determined on the base of criterion (5).

We state that the RMDL principle helps to select the best (of used set of representations) solution that minimizes the criterion $\|\varphi-\varphi_0\|$ (i.e. that has the best recognition rate on the patterns, which were not included in the learning sample). As the result, synthetic methods, which extend sets of used representations, should have better (or, at least, not worse) recognition rate than specific methods that compose this synthetic method. Some examples of extension of representation sets are presented below.

4. EXPERIMENTS

4.1. Selection of the best degree of polynomial discrimination function

Consider the case of two classes. Generalized discrimination functions are written in the form

$$\kappa(\mathbf{x} \mid \mathbf{w}) = \sum_{i=1}^{n} w_i y_i(\mathbf{x}) = \mathbf{w}Y(\mathbf{x}), \ \varphi(\mathbf{x}) = \begin{cases} 1, \kappa(\mathbf{x} \mid \mathbf{w}) < 0, \\ 2, \kappa(\mathbf{x} \mid \mathbf{w}) > 0, \end{cases} \quad (6)$$

where $\mathbf{w}$ is the parameter vector of the function $\kappa(\mathbf{x} \mid \mathbf{w})$, $y_i(\mathbf{x})$ is i-th generalized feature.

The discrimination function method utilizes predictive recognition approach, in which the RMDL criterion has the form $L(\{\mu_i\}_{i=1}^n, S\mathbf{w} \mid \{\mathbf{x}_i\}_{i=1}^n)$. Equation (6) describes algorithm S that reconstructs the class numbers μ on the base of the given parameters $\mathbf{w}$ and pattern $\mathbf{x}$.

Within some simplification assumptions one can obtain

$$L(\{\mu_i\}_{i=1}^n, S \mid \{\mathbf{x}_i\}_{i=1}^n) = \frac{M}{2}\log_2 n + \frac{n}{2}\log_2 \frac{\varepsilon^2(\mathbf{w})}{n}, \ \varepsilon^2(\mathbf{w}) = \sum_{i=1}^{n} [z_i - \mathbf{w}Y(\mathbf{x})]^2, \quad (7)$$

where M is the number of components in the parameter vector $\mathbf{w}$, and $z_i = -1$ if $\mu_i = a_1$, and $z_i = 1$ if $\mu_i = a_2$. Actually, this equation is not quite correct, but we have to omit technical details because of limited paper size. Nevertheless, the equation (7) can be used to select among the discrimination functions with different number of parameters. As an example, discrimination functions with different number of parameters were found for the set of patterns shown in fig. 1. Characteristics of these functions can be found in the table 1. It can be seen, that the solution with the minimum description length has also the best recognition rate on the new patterns ($\%_{test}$), which is not directly corresponds to the recognition rate on the learning sample ($\%_{learn}$).

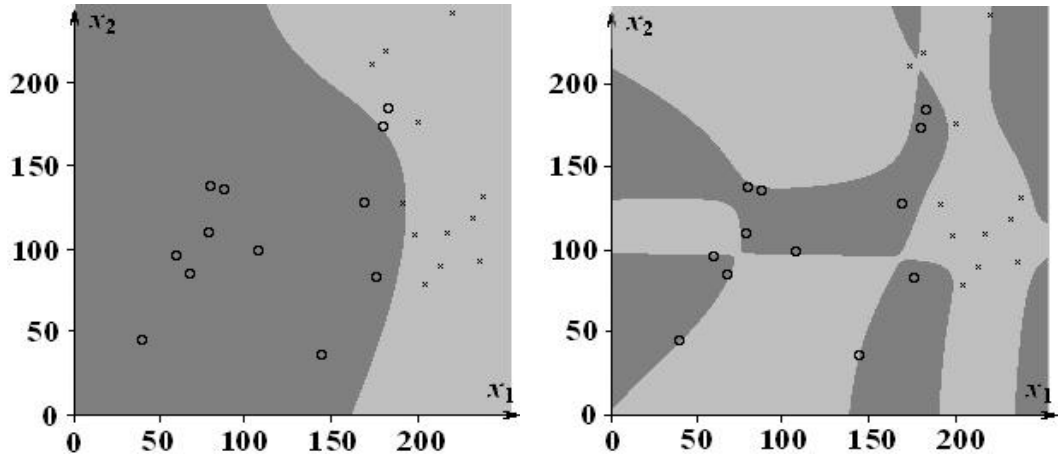

Figure 1. Discrimination functions with 4 and 25 parameters

Table 1. Comparison of discrimination functions of different complexity

№	M	$\%_{learn}$	$\%_{test}$	L, bits
1	4	16,7	6,4	26,9
2	9	12,5	8,5	33,7
3	16	0,0	23,0	36,7
4	25	0,0	41,0	55,0

These results can also be obtained with the use of classical MDL approach[1]. However, in addition, one can conclude on the base of the RMDL principle that the set of representations

can be extended with other basis functions, because they correspond to different algorithms *S***w**. Moreover, we wanted to show that existent methods utilize different pattern recognition problem settings, and this is the example of predictive recognition.

4.2. Selection among representations in the template matching method

One of the classic methods is the method of minimum distance or template matching classification. In the simplest case, each class μ is described by the etalon pattern (template) $\mathbf{y}_\mu$, and the decision function is specified in the form $\varphi(\mathbf{x}) = \arg\min_\mu \left\|\mathbf{x} - \mathbf{y}_\mu\right\|$.

One can propose the following representation, which corresponds to this method. Let pattern description within a representation consists of a class number μ and deviation $\delta = \mathbf{x} - \mathbf{y}_\mu$. Template $\mathbf{y}_\mu$ parametrize the representation: $\mathbf{w} = \{\mathbf{y}_\mu\}$. Program *S***w** of reconstruction of a pattern takes corresponding class number μ and deviation δ to produce $\mathbf{x} = \mathbf{y}_\mu + \delta$. However, different representations of deviations can be proposed.

Consider two families of representations $S^{(e)}\mathbf{w}$ and $S^{(g)}\mathbf{w}$, in the first of which distance from $\mathbf{x}$ to $\mathbf{y}_\mu$ is coded as the number of binary digits $\log_2\|\delta\|$ in the value $\|\delta\|$ followed by binary representation of value itself; and in the second case value $\|\delta\|$ is coded in accordance with the Gaussian distribution evaluated on the learning sample. For the task of inductive recognition one can obtain the following equations for these representations

$$L(D, S^{(e)}\{\mathbf{y}_\mu\}_{\mu=1}^d) = l(S^{(e)}) + l(\{\mathbf{y}_\mu\}_{\mu=1}^d) +$$

$$+ \sum_{\mu=1}^{d}\sum_{i=1}^{n_\mu}\left(l(\mu) + \log_2\left\|\mathbf{x}_{\mu,i} - \mathbf{y}_\mu\right\| + \log_2\log_2\left\|\mathbf{x}_{\mu,i} - \mathbf{y}_\mu\right\| + 1 + C\right),$$

$$L(D, S^{(g)}\{\mathbf{y}_\mu\sigma_\mu\}_{\mu=1}^d) = l(S^{(g)}) + l(\{\mathbf{y}_\mu\sigma_\mu\}_{\mu=1}^d) + \sum_{\mu=1}^{d}\sum_{i=1}^{n_\mu}\left(l(\mu) - \log_2 P\left(\left\|\mathbf{x}_{\mu,i} - \mathbf{y}_\mu\right\| \mid \sigma_\mu\right) + C\right),$$

where $\mathbf{x}_{\mu,i}$ is the *i*-th pattern that belongs to the μ-th class including n_μ patterns from the learning sample, *P* is Gaussian distribution with variance σ_μ^2, *C* is a constant.

Consider two learning samples (generated on the base of identical low) consisting of two classes presented in fig. 2. Templates found with the use of two families of representations differ.

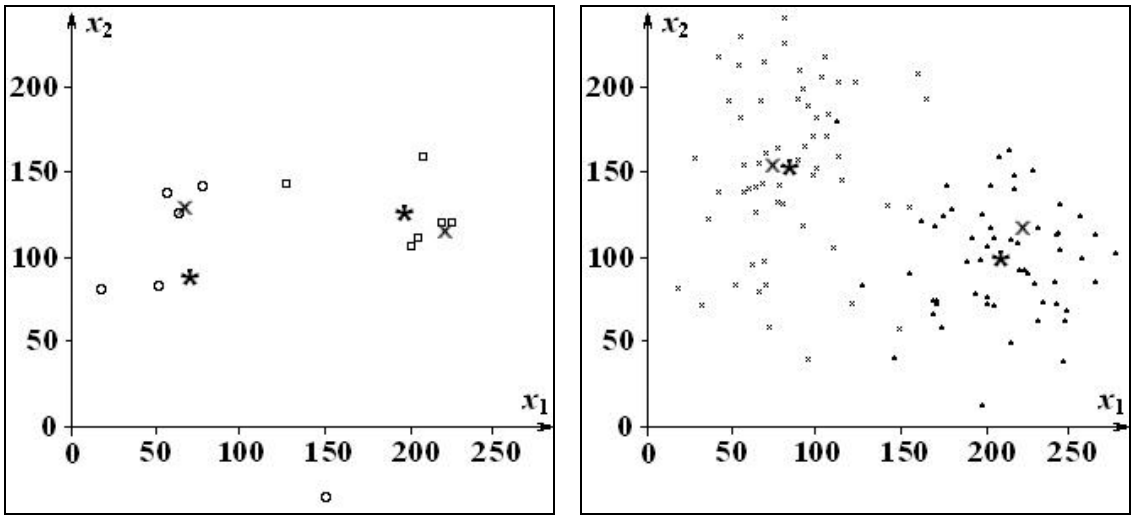

Figure 2. Example of templates (*, x) found with the use of different families of representation

As it seen in table 2, the best recognition rate corresponds to the template that can be chosen on the base of the RMDL criterion $L(D, S\mathbf{w})$.

Table 2. Comparison of $S^{(g)}$ and $S^{(e)}$

	Sample 1		Sample 2	
	$S^{(g)}$	$S^{(e)}$	$S^{(g)}$	$S^{(e)}$
$L(D,S\mathbf{w})$	92,84	**81,03**	**934,6**	1070,0
% errors	5,6	**3,9**	**2,9**	3,7

4.3. Extension of set of representation in Gaussian mixtures method

Classic approach for grouping is to use Gaussian mixtures to represent pattern distribution:

$$p(\mathbf{x} \mid \mathbf{w}) = \sum_{i=1}^{d} P_i \, p(\mathbf{x} \mid \mathbf{C}_i, y_i),$$

where P_i is the weight of the i-th component of mixture, p is the normal distribution with covariance matrix $\mathbf{C}_i$ and mean value $\mathbf{y}_i$, $\mathbf{w}$ is the total parameter vector.

The MDL principle helps to solve the problem of choosing the number of components d of the mixture[2]. The RMDL criterion can be reduced (with some simplifications) to the form

$$L(D, S\mathbf{w}) = l(S) + l(\mathbf{w}) - \sum_{i=1}^{n} \log_2 \, p(\mathbf{x}_i \mid \mathbf{w}).$$

The main point here is that we can try to extend the representation family. Let's additionally include some particular distributions: the one with diagonal covariance matrix $C_{i,i} = \sigma_i^2$, and another one with unitary matrix multiplied by variance $\mathbf{C} = \sigma^2 \mathbf{I}$. Families of representations with these simplified matrices differ from mixtures with full covariance matrices, thus there can be such the case, in which the best solution will belong to some of simplified families (at first sight, it can be surprising).

As an example, we generated learning and training samples of pattern in 5-dimentional feature space distributed within 3 clusters with coordinates. Three types of mixtures with different number of components were used. Corresponding description lengths are shown in the table 3.

Table 3. Lengths (in bits) of learning sample (n=24) described with different mixtures

Type	M=1	M=2	M=3	M=4
$\mathbf{C}$	**834**	855	855	-
$C_{i,i} = \sigma_i^2$	856	838	**817**	826
$\mathbf{C} = \sigma^2 \mathbf{I}$	859	857	826	**823**

As it can be seen, different number of components is chosen within different representation families, and the correct number can be chosen on the base of the RMDL criterion only when we extend Gaussian mixture method with some additional representation families.

5. CONCULSIONS

The representational MDL principle proposed in this paper fills the gap between theoretical and practical aspects of the MDL principle usage in pattern recognition. It is shown that tasks of pattern recognition and grouping can be reduced to the problem of selection of the best

representation, while the task of recognition of a single pattern is the problem of selection of the best description within the given representation. Practical methods of pattern recognition use very restricted families of representations, so constructed decision rule can be inadequate even with the entirely correct criterion. Thus, the problem of investigation of families of representations is the most urgent one.

We propose the notion of synthetic methods, in which automatic selection among particular families of representations is made. This selection can be performed on the base of the RMDL criterion. With some simple examples we showed that the use of synthetic methods helps to increate resulting recognition rate. However, some more powerful systems of patterns recognition on the base of synthetic methods are still to be developed.

It should be pointed out that the RMDL principle has different applications not only in pattern recognition, but also in computer vision as far as computer vision systems also deal with the mass problems. Moreover, some our preliminary experiments show that the RMDL principle can be more useful in computer vision, because it gives the base for a theory of selection of the best representations of images. But these results need further investigation.

REFERENCES

1. M. Sato, M. Kudo, J. Toyama, M. Shimbo, "Construction of a nonlinear discrimination function based on the MDL criterion," *1st Int. Workshop on Statistical Techniques in Pattern Recognition*, pp. 141–146 (1997).
2. H. Tenmoto, M. Kudo, M. Shimbo, "MDL-Based selection of the number of components in mixture models for pattern classification," in *Advances in Pattern Recognition*, number 1451 in *Lecture Notes in Computer Science*: Springer, pp. 831–836 (1998).
3. P. Vitanyi, M. Li, "Minimum description length induction, Bayesianism, and Kolmogorov complexity," *IEEE Trans. Information Theory*, vol. 46, no. 2, pp. 446–464 (2000).
4. R.J. Solomonoff, "Does algorithmic probability solve the problem of induction?" Oxbridge Research, P.O.B. 391887, Cambridge, Mass. 02139 (1997).
5. U. von Luxburg, O. Bousquet, B. Schölkopf, "A compression approach to support vector model selection," *J. Machine Learning Research*, vol. 5, pp. 293–323 (2004).
6. Solomonoff R., "The discovery of algorithmic probability", *J. Computer and System Sciences*, vol. 55, no. 1, pp. 73–88 (1997).