PATTERN RECOGNITION BY QUASI-LINGUISTIC TRANSLATION INTO ARTIFICIAL NOISE-RESISTANT LANGUAGE

A.N.Radchenko
Leningrad Polytechnic Institute,
Leningrad, USSR

Summary

A new approach to the recognition problem is considered, which does not require clustering of the sampling space into the appropriate number of regions.

Each sample mapped in multi-dimensional space is represented by a family of points (vectors), forming a certain configuration. Less compact location of the points leads to more simple and reliable pattern recognition. The input description is represented by the "text" formed by a Bet of binary pulse sequences. The recognition implies that the text fragments of different space-temporal structure should be related to one of the different classes. The recognition is carried out as an informational transformation of the input text into another one, invariant of any small variations in the space-temporal structure of the input text. Sufficiently different texts are translated into non-identical output texts. Repeated transformation of the input text by a number of translators is accompanied by an increasing "abstraction" (stabilization) of the output text relative to the input text variations. As a result, the output text structure can be considered as a list of classes, to which the input text fragments belong.

It is pointed out that the information compression characterizing the recognition process can not necessarily be carried out a priori but it can always be carried out as a consequence of the recognition process.

A correspondence is implied between the features of the "translation" process and some neuro-physiological phenomena.

Index terms: artificial intelligence, pattern recognition, clustering, information compressing.

Introduction

Methods for an automatic pattern recognition can be classified on the basis of:

1) forms of the input information representation,
2) the information coding methods,
3) methods for the coded signal processing.

The first item characterizes the physical information carrier and is not of principle importance, since the informational description of any physical

process (image, speech etc.) can be transduced to binary vectors and sequences.

The second item characterizes the relation between the physical process and the code. It is of importance that sufficient variations in the physical process must correspond to the sufficient variations in the output vector. As for small variations, they, as a rule, must correspond to small variations in the output vector. Such variations can often be neglected and it is not necessary to transfer them to the output vector. In other cases, on the contrary, the small variations must be emphasized by transforming them into larger output variations. Henceforth it will be enough to require that the coding does not distort sufficiently the metric relations which should be determined on the input and output signals of the encoder.

The third item is of principle importance. Many methods have been suggested to reduce the recognition problem to the clustering of a metric space into the domains the number of which is directly or indirectly related to the number of identified objects, i.e. to the construction of a deterministic or probabilistic "decoder" (1).

As an information transformation, recognition consists in the elimination of redundancy from the input vector. This process of information compression is facilitated in the earlier stages by detecting and eliminating less informative co-ordinates and by a reasonable amount of feature extracting from the original data. In any case, the information losses are accumulated during the information compression process, which sometimes leads to noticeable errors in the recognition results.

The pattern recognition method suggested in this work leads to lower information losses because the information compression (e.g. transition to a more laconic sign system) does not necessarily precede the recognition. At the same time, the information compression is quite feasible as a consequence of recognition in which the information losses are minimized.

The advantage of this work is the deliberate rejection of single-domain pattern localization in the sampling space and the rejection of the space clustering into domains, the number of which directly or indirectly depends on the number of classes. Each sample in this space is represented by a set of dislocated vectors (points). Patterns belonging to the same class are represented in the signal space by a set of domains the disposition of which to so-

me extent repeats and generalizes the disposition of points of a certain sampling.

A less compact disposition of these points and domains facilitates the solution of the sample belonging to the particular class. The larger is number of dislocated points contained in a single sample, the more reliable is their identification.

The recognition process is considered as a translation from one language into the other more economic one:

1. The input "text" is represented by a set of binary coded pulse sequences Fragments of this text are to be recogni zed differing in their space-temporal structure of the pulse position.

2. The classification results in the identical translation of the fragments which differ slightly. The translation of sufficiently differing fragments is also different.

3. Repeated transformation of the input text by a number of translators is followed by on increasing "abstraction" (stabilization) of output texts relative to the input text variations.

4. Stable output text fragments are decoded (recognition) or recoded by a more economic code (the information compression).

I. The essence of information processing

Let UB consider the original information which is encoded without losses and sufficient metric distortion into the binary pulse sequences, defined as a function of discrete time $t = 1,2,\dots,\gamma$

$$\begin{aligned} X_1 &= \{x_1(t)\} = x_1(1), x_1(2), \dots, x_1(\gamma) \\ X_2 &= \{x_2(t)\} = x_2(1), x_2(2), \dots, x_2(\gamma) \\ &\vdots \\ X_\kappa &= \{x_\kappa(t)\} = x_\kappa(1), x_\kappa(2), \dots, x_\kappa(\gamma) \end{aligned} \qquad /1/$$

where $x_i(t) = 0$ or 1 and γ is the sequence length.

Information /1/ will be called "the input text X". It is convenient to imagine the text /1/ as a description of pulse activity of the afferent nerve consisted of κ fibres, during the long time γ. To determine the metric the text /1/ can be represented in the form of sliding n-segments $Q(t)$:

$$2^n \gg \gamma \gg n \qquad /2/$$

$$X = \{Q(t)\} = \dots Q(t-1), Q(t), Q(t+1), \dots \qquad /3/$$

where $Q(t) = \begin{pmatrix} x_1(t-1), x_1(t-2), \dots, x_1(t-n) \\ x_2(t-1), x_2(t-2), \\ \vdots \\ x_\kappa(t-1), x_\kappa(t-2), \dots, x_\kappa(t-n) \end{pmatrix}$

is the "window" of the same "height" as the text /1/, but much more narrower than it.

For the convenience, the rows of $Q(t)$ can be rewritten in one line as an N-dimensional vector

$$N = Kn \qquad /4/$$

The notation /3/ enables UB to present the input text in the form of an N-dimensional cube apex sequence. This sequence given as a function of time, will be named "the signal trace''. The number of points on the signal trace under the accepted limitations /2/ almost equals the input text length.

$$y-n+1 \approx y$$

The distance between the points will be measured by the number of noncoincident elements of the vector *Aft)* . We shall be interested in many point segments of the signal trace $Q(t+1), Q(t+2), \ldots, Q(t+l)$ and in their belonging to particular classes. The conception of sample is not defined in the training process, since the segmentation problem arises in this case. It should be noted that not each segment of the signal trace belongs to one sample.

It should be pointed out that the Hamming distance between vectors and in the adjacent transition is sufficiently large. Actually, if /1/ represents the random realization with uniform and independent distribution of zeroes and units, then the distance probability p(d) in the adjacent transition is

$$p(d) = 2^{-N}\binom{N}{d}$$

It follows from the above formula, that on the average the "leap" is very large

$$E(d(Q(t), Q(t+1))) = \frac{1}{2}N \qquad /5/$$

The "leap" form is one of the essential properties of the signal trace.

Let us fix in N-cube the signal trace corresponding to a realization of the input text irrespective of whether the trace belongs to the particular classes. Using well-known taxonomic methods, the trace point set can be clustered into two parts in such a way that the boundary zone would be the largest. The separability criterion can take into account the number of short distance points belonging to different subsets, the values of their distances, etc. If the criterion for equidistantness of the separated subsets is chosen, then the separation surface can be constructed inside the boundary zone and the separation function can be defined

$$y(t) = F(Q(t)) \qquad /6/$$

where y *(t)* is "1", if the points are located on one of the surface sides, and "0", if the points are located on the opposite side.

The separation function /6/ allow the transformation of the text /1/ and other similar texts into a certain dif' ferent text I

$$Y=\{y(t)\}=y(1),y(2),\ldots,y(\gamma) \quad /7/$$

For the concordance of the text /1/ and /7/ lengths it is supposed that the first of them is added by n-1 zero signs which are not taken into account in tact numeration. If powers of the separated subsets are equal the leap pattern of the trace /3/ causes an approximately equal probability of transitions ($1\to1, 1\to0, 0\to1, 0\to0$) in the output text /7/, i.e. the absence of noticeable correlation between the adjacent symbols. It lays a foundation for obtaining high informative output reactions for the appropriate input.

In the case of long texts ($\gamma \gg n$) it is convenient to consider transformation as a translation from one language into the other one. The translation has the following properties:

1. Numerous slightly differing texts are transformed into an identical output text (the stabilization eifect). Sufficiently different texts correspond to the different translations. Indeed, a lot of small differences in the input text leads to proportional shifts of the signal trace points. If the removing point does not cross the separation surface, the output text does not change.

2. Recognition properties of the translation are not uniform in this process: the best stabilization is observed for the N-cube points located far from the surface, while the worst stabilization is observed for the points located near by the surface.

3. The effect of an increase in the variations is observed for the signal trace points located on the surface or in the vicinity of it. It is caused by the fact that the single distortion of the text/1/ leads to single shifts of *n* signal trace points. This feature should be taken into account in the separability criterion. Taking the different separation surfaces, one can ensure the wanted metric transformation of the input text.

4. A new text composed of "quotations" of the text used in the training process will be reproduced on the translator output in the form of associated fragments of the output text following the appropriately changed sequence. The limited quotation distortion do not affect the translation correctness.

5. Invariant output text fragments can be considered as a list of classes related to the appropriate quotations on the input. Variable output text fragments are associated with the quotation joints on the input. These output variations are grouped into bursts of maximum length $h-\frac{\rho}{K}-1$, where ρ is the halfwidth of the separation zone. If the

invariant fragments of the output text are considered as roots of words, then the variable ones could be considered as an adjustment of the grammar.

6. Strongly different input quotations are translated into different output "words". The shorter quotations can not be contained in the longer fragments. Thus, the sliding segments of the output text can be readily registered on the shift register by the decoder (recognition) or can be recoded into more economic code (information compression). On the other side, the above property can be considered as limitations on the minimum input text quotation length.

7. Taking the different separation surfaces, we determine one of the artificial languages for the translation. Clustering N -space into many regions and varying their designations, we can extend the variety of artificial languages and make an attempt to approximate one of them to the natural ones, e.g. to transform the speech into the teletype code.

8. It is possible to suggest that the clustering of N-space carried out after the long pithy training could lead to the creation of a translator ensuring a stabilizing translation (recognition) of the words, which are not used in the training.

2. Estimation of recognition efficiency in the worst indeterministic case.

Let us suppose that the input text / 1 / is represented by a set of binary pulse sequences, in which zeroes and units are distributed independently and uniformly. It is necessary for the recognition translation of the text to know if it is possible to separate the signal trace points into two subsets in such a way that the points would not move from one subset to the other, as a result of single shifts of the points. This condition is deliberately observed if the subset points are distant from the separation surface at least by one of the co-ordinates.

A calculation will be given below of a posteriori division of the random N-cube points by the zone consisting of random points.

If these separated subset are contained in two Hamming's spheres with their centers located at the princip diagonal, then the number M of possible separation zones is equal to the number of complementary vector pairs of the N-cube points. Then we get

$$m = 2^{N-1}$$

Each separating zone contains b points

$$b = 2^N - 2\sum_{i \le r} \binom{N}{i}$$

where r is the sphere radius.

Let each of γ tests result in elimination of the zones containing $Q(t)$ from the separation zone set. If even one zone will remain at the end of the tests, then the desired division is possible. The division probability should be determined.

The general separation zone set has a symmetric location in N-cube according to the zone formation law. Each point of N-cube belongs simultaneously to the different separation zones. The number of the zones containing one point is

$$\Delta = \frac{mb}{2^N} = \frac{b}{2}$$

Suppose that it is possible to select γ of m variants in such a way that their zones form a single-layer packing containing all N-cube points, and that the general set of /*! zones can be represented by Δ-layer packing. It is evident that

$$\gamma = \frac{2^N}{b}$$

Actually such a division can be carried out only approximately. For the particular layer, the probability p_1 that all separation zones will contain the points $Q(t)$ in γ independent tests is given by (3)

$$p_1 = \binom{\gamma-1}{\gamma-1} : \binom{\gamma+\gamma-1}{\gamma-1}$$

Hence the probability that at least one neutral separation zone can be founded among Δ layers is

$$Q = 1 - p_1^{\Delta} = 1 - \left(\frac{\alpha-1}{\alpha+1}\right)^{b/4}\left[\frac{\alpha-1}{\alpha+1}\left(1-\frac{1}{\alpha^2}\right)^{\alpha}\right]^{2^{N-1}} \quad /8/$$

where $\alpha = \gamma b 2^{-N}$

In the case of $\alpha \gg 1$ it follows from formula /8/ that

$$Q \approx 1 - \exp\left\{-\frac{2^{N-1}}{\alpha} - \left(2^N + \frac{b}{2}\right) \operatorname{Arcth}\alpha\right\} \approx$$

$$\approx 1 - e^{\frac{3}{\alpha}2^{N-1}} \quad /9/$$

The second member vanishes with increasing N. Hence the desired division of the signal trace points into subsets can be readily found. In connection with the suggestion of N-cube packing separability, the results /8/ and /9/ should be considered as the upper limit estimation The lower limit estimation can be obtained on the assumption that each layer of N-cube packing contains only one of m neutral zones.

The probability for the point $Q(t)$ to be contained in one of the fixed separation zones in a single test is

$$q = b 2^{-N}$$

The current value of the remaining zone number is designed by $m(t)$. Then the mean number of the single ube point packing by the separation zones is

$$\Delta(t) = 2^{-N} m(t) b = m(t) q$$

The expected value of the remaining zone number after the next test is $m(t+1)$

$$m(t+1) = m(t) - \Delta(t) = m(t) p$$

where $p = 1 - q$

Then the expected value of the number of the zones remaining after the training can be readily obtained:

$$m(\gamma) = m(0) p^{\gamma}$$

The probability that each particular zone will remain after the training is p^{γ}. The probability that all zones will be removed is

$$P > (1-p^{\gamma})^{m}$$

The probability that at least one separation zone will remain is given by

$$R = 1-[1-(1-b2^{-N})^{\gamma}]^{2^{N-1}} \qquad /10/$$

Suppose that the condition

$$(N-1)\ln 2 - \frac{\gamma b}{2^{N-b}} = \ln \omega \qquad /11/$$

is satisfied, where ω is a real number Then

$$R > 1-e^{-\omega} \qquad /12/$$

It follows from /11/ that to obtain $R \to 1$ the linear function of N must be the upper limit of the value $\alpha = \gamma b 2^{-N}$. Hence the desirable separation of the signal trace points into subsets can be found. The larger is the easier this separation can be carried out.

3. The worst deterministic case.

Much worse than the random location of the signal trace points in N-cube is a set of locations so determined that all γ points occupy uniformly the N-cube volume, being separated by equal distances $d = 2r+1$

Then we have

$$2^{N} \geq \gamma \sum_{i \leq r} \binom{N}{i} \qquad /13/$$

It is necessary for stabilizing translation to ensure $r \geq 1$ for any pair of points. One can find a certain N so that this condition will be observed.

Using the Shannon's approximation (2)

$$\sum_{i \leq r} \binom{N}{i} \approx 2^{NH(\frac{r}{N})}$$

which is valid for $r/N < \frac{1}{2}$; $r, N \to \infty$ we get from /13/

$$N(1-H(\tfrac{r}{N})) \geq \log_2 \gamma \qquad /14/$$

where $H(\frac{r}{N}) = -\frac{r}{N}\log_2\frac{r}{N} - \frac{N-r}{N}\log_2\frac{N-r}{N}$

The larger N is the easier the eq./14/ is satisfied. Hence there are no difficulties in the separation even in this case.

Conclusions.

According to the above considerations the following conclusions can be drawn:

1. Correspondence between the number of distinctive patterns and that of the appropriate separable domains of the sign space is not necessary for the pattern classification.

2. The transformation suggested in this work does not require a priori solution of the input text segmentation problem.

3. To improve the classification the dimensions of the signal space should not be reduced. Hence:

a) transition from the initial co-ordinates to a more economic sign system is not necessary.

b) less informative co-ordinates should not be left out.

4. The suggested method enables ue to carry out the information compression

within any limits given in advance as a consequence of the recognition,

5. The suggested method has some features of its neuro-physiological prototype, the brain, in particular, the distributivity phenomenon. The conception of recognition translation is the linguistic treatment of the conditioned reflex(4).

It should be noted that the segmentation problem solution in this case is closely related to the problem of shift, angle and size invariance in the pattern recognition. The solution of those problems is difficult, since the samples belonging to the same class have large Hamming's distances. At the same time, it is usually impossible to find any metrics which would not be connected with abrupt leap variation in the distance in one-step transition. One can see that this "difficulty" has been used in this work as a natural basis for the separation and construction of noise-re sistant (stabilizing) translation into an artificial language.

Hence one should try to obtain the "leap" signal trace by the appropriate choice of the quantization step, metrics, scanning etc. The seeming increase in complication of the recognition process due to the multiple representation of samples in an /r-cube is not the case indeed, since the ube division can be made by the simplest methods. The lack of a priori limitations for the number of separated domains also facilitates the recognition problem in comparison with other available methods. Current clustering techniques and in particular, tazonomic methods are entirely acceptable for the solution of the recognition problem by the methods mentioned above.

References

1. V.I.Vasiliev, "Recognition systems", Naukova dumka, Kiev, 1969.
2. W.Peterson, "Error-correcting codes", Mir, Moscow, 1964.
3. W.Feller, "Introduction to the theory of probability and its applications", Mir, Moscow, 1967.
4. A.N.Radchenko "Analytical study of neuro-physiological processes of information recording and reproduction on the speculate model", in "Cybernetical aspects in the study of the brain action", Nauka, Moscow, 1970.