Вестник Ленинград. Унив.
1972, № 13

Vestnik Leningrad Univ. Math.
Vol. 5 (1978)

ON THE PROBLEM OF SYNTHESIZING SELF-LEARNING RECOGNITION SYSTEMS

UDC 519.95

A. L. FRADKOV

Abstract. In this paper the problem of a self-teaching process of pattern recognition is treated as the reconstruction by an automaton of a certain a priori classification of the input objects. These are regarded as a sequence of points in Euclidean n-space forming a sample in an n-dimensional universe with an unknown continuous density distribution $\pi(x)$. Two algorithms are described which yield classifications of the sample arbitrarily close to the a priori one, with probability tending to 1 as the size of the sample grows indefinitely. In the second algorithm only part of the input sample is correctly indexed, but the length of the correctly indexed part is proportional to the size of the sample. The results of computer experiments with both algorithms are described.

Bibliography: 6 titles.

One of the urgent problems of engineering cybernetics is the analysis and synthesis of recognition systems. Recognition systems are automata that divide the set of input events into several subsets, called classes, i.e. that classify input events. Of particular interest is the synthesis of learning and self-learning systems, in which all the information necessary to design a decision rule is derived by selecting from the set of input events a so-called learning sequence. Here a self-learning system is understood to mean a system without any information about the distribution of the images among the classes. The system itself must determine whether the images of a learning sequence belong to some class or another, and must index the learning sequence.

Let us consider two possible approaches to the design of a self-learning system. Suppose that images (objects), each of which is specified by an ordered set of n numerals, are fed to the input of the system. Each image will be regarded as a point of the Euclidean space R^n. A learning sequence can then be considered as a sample from an n-dimensional universe with a certain distribution density $\pi(x)$. The number of classification classes is assumed to be given. The design of a self-learning system (self-learning algorithm) can be stated naturally as the task of creating an algorithm that generates a "good" indexing of the learning sequence. However, the problem remains meaningless until the term "good" is made precise.

1980 *mathematics subject classification.* Primary 68G10.

One possible refinement is to introduce some classification quality criterion. Then the task becomes that of constructing an algorithm of searching for a classification that is optimal in the sense of this criterion. Since a classification can usually be specified by a set of separating functions (hypersurfaces in the image space), the problem reduces to the design of algorithms for the numerical solution of a variational problem. Such an approach has been developed in [1]–[4]. A typical example of a quality functional is the ratio of the mean-square distance between points of different classes to the mean-square distance between points of the same class [3]. Extremalization of this and other similar criteria makes it possible to obtain classifications in which the classes are fairly "compact" and "distant" sets. The choice of a criterion is to a certain extent arbitrary and is dictated by considerations of naturalness and convenience for numerical extremalization.

A second approach, which is developed, in particular, in the present paper, is to define as "good" a classification close (in some sense) to a previously given classification which is not known a priori to the computer. The problem can then be stated as the design of an algorithm that reconstructs this a priori classification. If the form of the a priori classification is arbitrary, the task of reconstructing it is meaningless. An a priori classification must therefore satisfy certain conditions.

It is of interest to solve the problem under the weakest possible constraints on the form of the a priori classification and the function $\pi(x)$. Suppose the classes $A_1, \ldots, A_k$ of the a priori classification are disjoint bounded open connected sets in R^n.

We now describe the following self-learning algorithm (Algorithm 1).

DEFINITION 1. We specify a number $R > 0$. A set M of points in Euclidean space is said to be *R-connected* if for every pair of points in M there exists a polygonal line connecting them, with vertices at points in M, such that the length of each link is at most R.

DEFINITION 2. Suppose $M \subset R^n$ and $x \in M$. The greatest R-connected subset of M containing the point x is called the *R-connected component* of x.

ALGORITHM 1. Let $x_1, \ldots, x_N$ be a learning sequence ($x_i \in R^n$), and let $R > 0$. We construct the R-connected component of x_1. Then we take an arbitrary point from the learning sequence outside this component, say x_{i_0}. We construct the R-connected component of x_{i_0}, and continue until the entire learning sequence has been exhausted. At each step we construct the R-connected component of a point that does not occur in any previously constructed component.

It can be easily proved that as a result a learning sequence splits into

One possible refinement is to introduce some classification quality criterion. Then the task becomes that of constructing an algorithm of searching for a classification that is optimal in the sense of this criterion. Since a classification can usually be specified by a set of separating functions (hypersurfaces in the image space), the problem reduces to the design of algorithms for the numerical solution of a variational problem. Such an approach has been developed in [1]–[4]. A typical example of a quality functional is the ratio of the mean-square distance between points of different classes to the mean-square distance between points of the same class [3]. Extremalization of this and other similar criteria makes it possible to obtain classifications in which the classes are fairly "compact" and "distant" sets. The choice of a criterion is to a certain extent arbitrary and is dictated by considerations of naturalness and convenience for numerical extremalization.

A second approach, which is developed, in particular, in the present paper, is to define as "good" a classification close (in some sense) to a previously given classification which is not known a priori to the computer. The problem can then be stated as the design of an algorithm that reconstructs this a priori classification. If the form of the a priori classification is arbitrary, the task of reconstructing it is meaningless. An a priori classification must therefore satisfy certain conditions.

It is of interest to solve the problem under the weakest possible constraints on the form of the a priori classification and the function $\pi(x)$. Suppose the classes $A_1, \ldots, A_k$ of the a priori classification are disjoint bounded open connected sets in R^n.

We now describe the following self-learning algorithm (Algorithm 1).

DEFINITION 1. We specify a number $R > 0$. A set M of points in Euclidean space is said to be *R-connected* if for every pair of points in M there exists a polygonal line connecting them, with vertices at points in M, such that the length of each link is at most R.

DEFINITION 2. Suppose $M \subset R^n$ and $x \in M$. The greatest R-connected subset of M containing the point x is called the *R-connected component* of x.

ALGORITHM 1. Let $x_1, \ldots, x_N$ be a learning sequence ($x_i \in R^n$), and let $R > 0$. We construct the R-connected component of x_1. Then we take an arbitrary point from the learning sequence outside this component, say x_{i_0}. We construct the R-connected component of x_{i_0}, and continue until the entire learning sequence has been exhausted. At each step we construct the R-connected component of a point that does not occur in any previously constructed component.

It can be easily proved that as a result a learning sequence splits into

In actual problems, however, $\pi(x)$ by no means always satisfies Conditions 1 and 2. It usually does not satisfy Condition 2, i.e. "noise points" blurring the interval between classes occur with nonzero probability. It is necessary to perfect the algorithm in such a way that it could index the learning sequence correctly under weaker constraints on the form of $\pi(x)$.

Let us make these remarks more precise. Suppose that a density function $\pi(x)$ and k disjoint bounded connected and open sets $A_1, \ldots, A_k$, $A = \bigcup_1^k A_j$, are defined in R^n. Let $\pi(x) > 0$ for $x \in A$. We set

$$\alpha_0 = \sup_{x \in R_n \setminus A} \pi(x).$$

CONDITION 3. $\pi(x) > \alpha_0$ for any point $x \in A$.

Condition 3 expresses the natural requirement that the probability density for noise points to appear must not be too large; namely, that it is less than the probability density for points to belong to classes. The constraints imposed on the form of $\pi(x)$ by Condition 3 are weaker than the constraints imposed by Condition 2.

We now introduce some notation. By $P(M)$ we denote the probability that a point belongs to some measurable set M. Then

$$P(M) = \int_M \pi(x)\, dx. \tag{1}$$

By $S_R(x)$ we denote an n-dimensional ball of radius R with center at x, by V_R its volume.

DEFINITION 3. The function $\pi_R(x)$ defined for every point $x \in R^n(x)$ by

$$\pi_R(x) = \frac{P(S_R(x))}{V_R} \tag{2}$$

is called the *R-density function.* The R-density at x is the mean value of the density $\pi(x)$ in $S_R(x)$.

Note the following properties of $\pi_R(x)$:

1) $\inf_{y \in S_R(x)} \pi(y) \leqslant \pi_R(x) \leqslant \sup_{y \in S_R(x)} \pi(y)$;

2) $\pi_R(x) \leqslant 1/V_R$;

3) $\lim_{R \to 0} \pi_R(x) = \pi(x)$ if $\pi(x)$ is continuous;

4) $\pi_R(x)$ is continuous for any R if $\pi(x)$ is bounded.

Thus, for small R, $\pi_R(x)$ can be regarded as a bound on the continuous density function $\pi(x)$.

Suppose that from the universe a sample of size N is selected with density function $\pi(x)$. We let d_i denote the number of points of the sample occurring in $S_R(x_i)$, divided by the sample size. The number d_i is a bound for the

quantity

$$P\{\lim_{N\to\infty} d_i = \pi_R(x_i)\, V_R\} = 1. \tag{3}$$

Let us also introduce the notation

$$M_R(\alpha) = \{x : \pi_R(x) \leqslant \alpha\}, \tag{4}$$

$$C(R, \alpha) = P(M_R(\alpha)). \tag{5}$$

We now describe

ALGORITHM 2. A radius $R > 0$ and threshold $\alpha > 0$ are fixed. Points x_i with $d_i \leqslant \alpha V_R$ are eliminated from the learning sequence. Algorithm 1 is then applied to the remaining points.

The use of Algorithm 2 is justified by the following theorem.

THEOREM. 1. *For any positive numbers R, α and ϵ, the probability that "Algorithm 2 eliminates at most $(C(R, \alpha) + \epsilon)N$ points from the learning sequence" tends to 1 as $N \to \infty$.*

2. *For any $\epsilon > 0$ there exist positive numbers R_0 and α_0 such that for all positive $R < R_0$, the probability that $N_1/N_0 > 1 - \epsilon$ tends to 1 as $N \to \infty$. Here N_0 is the number of points of the learning sequence left after the elimination and N_1 is the number of correctly indexed points.*

The meaning of the theorem is that Algorithm 2 eliminates at most cN points, $0 < c < 1$, from the learning sequence and indexes the remaining points in such a way that the probability for the correctly indexed fraction of the sequence to be greater than $1/(1 - \epsilon)$ for any $\epsilon > 0$ tends to 1 as $n \to \infty$.

PROOF. By (3), the points eliminated by Algorithm 2 form a sample from the set $M_R(\alpha)$ with probability tending to 1 as $N \to \infty$. If there are l such points, then

$$\lim_{N\to\infty} \frac{l}{N} = P(M_R(\alpha)) = C(R, \alpha) \tag{6}$$

with probability 1. Consequently, for any $\epsilon > 0$ the probability for

$$C(R, \alpha) - \varepsilon < \frac{l}{N} < C(R, \alpha) + \varepsilon \tag{7}$$

tends to 1 as $N \to \infty$, which proves the first assertion of the theorem.

We introduce some more notation. Let M be a subset of R^n. We denote by M_{+R} the set of points in R^n whose distance from points in M is at most R; by M_{-R} the set of points in M at a distance at least R from the boundary of M; and by $\overline{M}_R(\alpha)$ the set of points at which the R-density is greater than α:

$$M_{+R} = \{x : \rho(x, M) \leqslant R\}, \tag{8}$$

$$M_{-R} = \{x : \rho(x, R^n \setminus M) \geqslant R\}, \tag{9}$$

$$\overline{M}_R(\alpha) = \{x : \pi_R(x) > \alpha\}. \tag{10}$$

If Condition 3 is satisfied, then, for any $R > 0$,

$$A_{-R} \subset M_R(\alpha_0) \subset A_{+R}. \tag{11}$$

We claim that for any $\epsilon > 0$ there is an $R > 0$ such that

$$\text{the classes } (A_i)_{+R} \text{ satisfy Condition 1;} \tag{12}$$

$$P((A_i)_{-R}) > 0, \quad 1 \leqslant i \leqslant k; \tag{13}$$

$$P(A_{+R} \setminus A_{-R}) < \varepsilon. \tag{14}$$

Now (12) is true if R is chosen less than $\delta/2$. where δ is the least distance between the classes, in accordance with Condition 1; and for (13) to hold it is sufficient that the sets $(A_i)_{-R}$ are nonempty and contain interior points, and this is true for sufficiently small R, since the A_i are open sets. For small R (14) is also satisfied, since $P(A_{+R} \setminus A_{-R}) \longrightarrow 0$ as $R \longrightarrow 0$ (we assume that $\pi(x)$ is bounded). We select a value for R satisfying (12)–(14) and set $\gamma_0 = P(M_R(\alpha_0)) > 0$, where α_0 is the number in Condition 3.

We now apply Algorithm 1 to the points of the learning sequence for which $d_i > \alpha_0 V_R$. As was proved above, these points constitute a sample from $M_R(\alpha_0)$ with probability tending to 1 as $N \longrightarrow \infty$. If there are l such points in the sample, then $l/N \longrightarrow \gamma_0$. It is clear from (11) that the set $\overline{M}_R(\alpha_0)$ splits into k disjoint subsets:

$$M = \bigcup_{i=1}^{k} M_i, \quad \text{while} \quad (A_i)_{-R} \subset M_i \subset (A_i)_{+R}.$$

The sets M_i satisfy Conditions 1 and 2. Consequently, Algorithm 1 divides the chosen points into k groups with probability tending to 1 as $N \longrightarrow \infty$, where the ith group is a sample from M_i (also with probability tending to 1 as $N \longrightarrow \infty$).

The set $\overline{M}_R(\alpha_0)$ does not, in general, coincide with A, and the algorithm incorrectly indexes points in $\overline{M}_R(\alpha_0)$ that do not belong to A. However, points in A_{-R} are correctly indexed. The probability that a point has been incorrectly indexed can be made arbitrarily small by the choice of ϵ.

Since

$$P(\overline{M}_R(\alpha_0) \setminus A_{-R}) < P(A_{+R} \setminus A_{-R}) < \varepsilon,$$

at least $(\gamma_0 - \epsilon)N$ points will be correctly indexed with probability tending to 1 as $N \longrightarrow \infty$. But $\gamma_0 = 1 - C(R, \alpha_0)$, and we may conclude from (7) that the

probability for

$$\frac{N_1}{N_2} > \frac{\gamma_0 - \varepsilon}{\gamma_0 + \varepsilon}$$

to hold tends to 1 as $N \longrightarrow \infty$, for any $\epsilon > 0$, which proves the theorem.

REMARK 1. Since Conditions 1 and 3 are together weaker than Conditions 1 and 2, the statement of the theorem is also weaker: correct indexing is ensured not of the entire learning sequence, but only of part of it. The length of the correctly indexed segment is the greater, the lesser is the probability density of noise points to appear.

REMARK 2. The set of classifications of a universe with density function $\pi(x)$ can naturally be made into a metric space by defining the distance between two classifications as the probability of the set of points by which they differ:

$$\rho = \sum_{i=1}^{k} P(A_i \Delta B_i), \tag{15}$$

where B_i are the classes of the second classification.

To make the distance so defined independent of the order in which the classes are numbered, we take the least ρ for all possible methods of numbering. The theorem can then be stated in the following way: *Suppose the a priori classification satisfies Conditions* 1 *and* 3 *and a sample of size N enters the input. Algorithm* 2 *then selects from it a subsequence and indexes it in such a way that it gives rise to a classification arbitrarily close to the a priori classification with probability tending to* 1 *as* $N \longrightarrow \infty$.

The application of Algorithm 2 is illustrated by the following experiment. A learning sequence consists of 74 images (the results of measuring the species *Ch. concina, Ch. heikertingeri,* and *Ch. heptapotamica* of flea beetles [6]). There are $k = 3$ classes. When $R = 13.0$, ten points with $d_i \leqslant 2$ are eliminated from the sequence. Algorithm 1 is applied to the remaining points and for this value of R splits the remaining learning sequence into three R-connected groups. All the points of each group belong to the same species, i.e. the indexing is correct.

Experiments have demonstrated that the range of R over which estimates of R-density were calculated and a fairly small variation of the threshold α_0 affect the result of the algorithm only insignificantly.

The author wishes to express his appreciation to V. I. Ružanskiĭ and V. A. Jakubovič for a number of valuable remarks and for their interest in the work.

Received 21/MAY/1970

BIBLIOGRAPHY

1. M. I. Šlezinger, *On spontaneous image recognition,* Reading Automata and Image Recognition, "Naukova Dumka", Kiev, 1965, pp. 38–45. (Russian)

2. È. M. Braverman, *The method of potential functions in the problem of training machines to recognize patterns without a teacher,* Avtomat. i Telemeh. 1966, no. 10, 100–121; English transl. in Automat. Remote Control 27 (1966).

3. V. A. Laptev and A. V. Milen'kiĭ, *Pattern recognition in the self-training mode,* Izv. Akad. Nauk SSSR Tehn. Kibernet. 1966, no. 6, 110–115; English transl. in Engrg. Cybernetics 1966.

4. Ja. Z. Cypkin and G. K. Kel'mans, *Recursive self-training algorithms,* Izv. Akad. Nauk SSSR Tehn. Kibernet. 1967, no. 5, 78–87; English transl. in Engrg. Cybernetics 1967.

5. V. I. Ružanskiĭ, *Certain "non-incentive" learning algorithms for recognition automata,* Vyčisl. Tehn. i Voprosy Programmirovanija [Leningrad. Gos. Univ.] vyp. 4 (1965), 75–83. (Russian)

6. G. M. Genkina, O. M. Kalinin and Š. D. Fljate, *Discriminant analysis for distinguishing three or more sets, with an application to the taxonomy of flea beetles of genus Chaetocnema,* Problemy Kibernet. Vyp.18 (1967), 147–154; English transl. in Systems Theory Res. 18 (1967).

Translated by R. H. SILVERMAN