The Application of Genetic Algoritm for Training “Without a Teacher”

Valeriy FRATAVCHAN
Yurii Fedkovich Chernivtsi National University
str.Universitetsika.1,Chernivtsi
vgfrat@mail.ru

Abstract — **The algorithm of determination of reference patterns for classification in the conditions of training «without a teacher» is described in this paper. The case is considered when patterns are set by n – measured vectors of numerical stochastic signs. For finding the coordinates of reference vectors it is offered to use Genetic Algorithm.**

Index Terms — **Recognition, Training, Self-training, Vector of Signs, Space of Signs, Optimal Estimations, Genetic Algorithm, Genetic Operations**

I. INTRODUCTION

The most widespread method of recognition system adaptation to a concrete set of patterns is a training method «with a teacher». While using this method at the initial stage of the recognition system work a highly-competent user indicates number of classes, creates training sequence of samples and specifies to the system, what class possesses each sample of training sequence in an interactive mode. As a result of statistical processing of samples the system defines working parameters of each class or creates the generalized reference samples for each class.

Sometimes there are situations when it is impossible to apply an interactive method of "prompting" or a "teacher" cannot define unambiguously to which class this or that image should be referred to. In such cases the training method «without a teacher» is used. It is possible to distinguish two forms of this method. In the first case recognition system is specified and the number of classes and training sequence is created. The system must divide the training sequence into the subsets belonging to different classes and define parameters of these classes independently [1]. In the second case the number of classes is not even known preliminarily. Only succession of samples of training sequence is given to a system. The recognition system defines the number of classes, divides the training sequence into subsets and calculates the parameters of classes independently.

This paper proposes the algorithm of training «without a teacher» for a case of classification by a method of comparison with the template in which the number of classes is known, and patterns are described by n – measured vectors of numerical stochastic signs.

II. MATHEMATICAL PROBLEM STATEMENT

Let members of some general set Ω be described by the vectors of signs

$$X = (x_1, x_2, ..., x_n), x_i \geq 0, x_i \in R,\ i = 1,...,n.$$

Let $S = \{X^1, X^2, ..., X^N\}$ are vectors of signs of some casual sample from the set Ω.

It is known that the set Ω consists of K subsets – classes:

$$\Omega = \Omega^1 \cup \Omega^2 \cup ... \cup \Omega^K,$$
$$\Omega^i \cap \Omega^j = 0, \quad i \neq j, \quad i,j = 1,...,K. \tag{1}$$

We assume that sample S is big enough, and some members of each class have been got to it. It is necessary to divide the sample S into subsets (S splitting):

$$S = S^1 \cup S^2 \cup ... \cup S^K,$$
$$S^i \cap S^j = 0, \quad i \neq j, \quad i,j = 1,...,K. \tag{2}$$

Vectors of signs of subsets members $\Omega^i,\ i = 1,...,K$ must get to each subset S^i, $i = 1,...,K$ accordingly.

For splitting of the set S into subsets the optimizing task is used

$$F(S) \to \min, \tag{3}$$

where $F(S)$ is a certain estimation of splitting.

This estimation is created on the basis of the following considerations. Let probabilities of the classes are equal:

$$P(\Omega^1) = P(\Omega^2) = ... = P(\Omega^K).$$

Let in attribute space to each sign classes have identical mean-square deviations from the corresponding own average values. It allows to use for classification the formulae of distances in n - measured space:

a) Euclid’s measure:

$$d(X^1, X^2) = \sqrt{\sum_{i=1}^{n} (x_i^1 - x_i^2)^2},\ X^1, X^2 \in R^n;$$

b) Manhattan measure:

$$d(X^1, X^2) = \frac{1}{n}\sum_{i=1}^{n} \left| x_i^1 - x_i^2 \right|,\ X^1, X^2 \in R^n;$$

c) Chebyshev’s measure:

$$d(X^1, X^2) = \max_{i=1,...,n}(\left| x_i^1 - x_i^2 \right|),\ X^1, X^2 \in R^n.$$

Since the method of comparison with the standard for the classification is used, we assume that the reference vector of signs is chosen for the each class:

$$E = \{E^1, E^2, ..., E^K\},\ E^1 \to \Omega^1, ..., E^K \to \Omega^K.$$

In that case it is possible to present a splitting estimation as:

$$F(S) = \sum_{i=1}^{K} \sum_{j=1}^{K_i} d\left(E^i, X^{i,j}\right) \tag{4}$$

where

K^i is a quantity of members of the sample for an i - class;

E^i is the class of an i standard;

$X^{i,j}$ is a vector of signs of a j - sample of training sequence for an i -class.

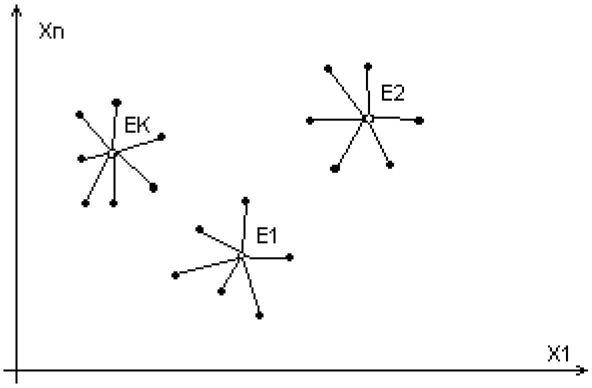

Figure 1.

Training process is reduced to the search of an optimal, from the point of view of an estimation $F(S)$, arrangement of signs K of standards $E^1, E^2, ..., E^K$ in a space [2].

Such tasks are efficiently solved by means of Genetic Algorithm.

III. STATEMENT OF THE OPTIMIZING PROBLEM FOR GENETIC ALGORITHM

Analyzing vectors of signs of the sample, the vector of the maximum values of signs is calculated:

$$M = (m_1, m_2, ..., m_n),$$
$$m_i = \max_{j=1,...,N}(x_i^j). \tag{5}$$

The problem is considered [3]:

$$F(S;E) = F(S;E^1;E^2;...;E^K) =$$
$$= F(S;e_1^1, e_2^1, ..., e_n^1;...;e_1^K, e_2^K, ..., e_n^K) \to \min \tag{6}$$

assuming $e_i^j \le m_i,\ i = 1,...,n;\ j = 1,...,K$.

For conducting of a splitting under the chosen system of standards it is necessary to classify the samples of training sequence, that is to define, what class possesses each training image from the point of view of its comparison with the specified set of standards:

$$k = \arg\min_j (d(X, E^j)) \tag{7}$$

where

X is a current member of training sequence;

E^j – is a reference pattern of a j - class;

k is a received number of a class.

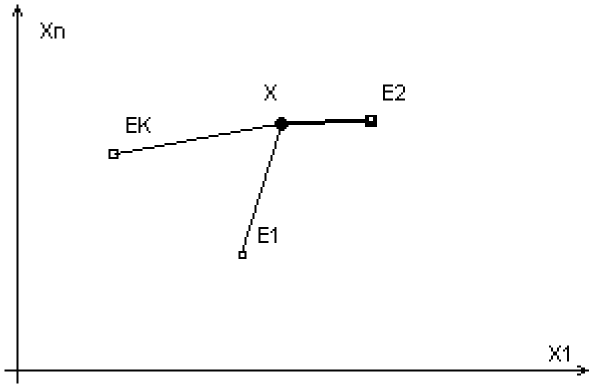

Figure 2.

Then objective function corresponds to a splitting estimation S at standards E:

$$F(S,E) = \sum_{i=1}^{K} \sum_{j=1}^{K_i} d\left(E^i, X^{i,j}\right). \tag{8}$$

IV. THE ORGANIZATION OF GENETIC ALGORITHM

For embedding of Genetic Algorithm it is necessary to define preliminarily the maximum values of signs for each sign and an approximation step. The maximum values of signs are included in mathematical model of an optimizing problem (5). The approximation step is calculated by the formula:

$$h_l = \min_{i \ne j; h>0} (\left|x_l^i - x_l^j\right|);\ i, j = 1,...,N, l = 1,...,n \tag{9}$$

h_l is a step of approximation of l -sign.

For completeness of the model we consider that each sign can have own scale parameters for representation in Genetic Algorithm.

Considering scale parameters $\{m_l, h_l\}, l = 1,...,n$, the variety of standards sets – initial "population" is created by means of the random numbers generator:

$$\{E(i)\}, i = 1,...,R$$

where

i is a number of a standards set;

R is the size of population;

$E(i) = \{E^1(i), E^2(i), ..., E^K(i)\}$ - i- variant of arrangement of reference vectors in attribute space (gene).

For each set of standards $E(i)$, splitting of training sequence $S(i)$ is carried out under the formula (7) ; then the splitting and feature set estimation is calculated under the formula (8)

$$F(i) = F(S(i), E(i)).$$

Variants (genomes) $\{E(i), F(i)\}, i = 1,\dots,R$ are put in order on estimations decrease of $F(i)$.

Consequent "generations" are calculated under the classical scheme [4]:

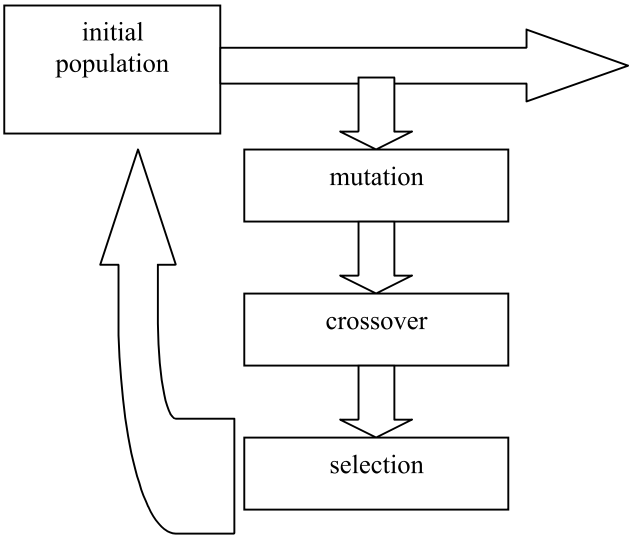

Figure 3.

Members for genetic operations during the construction of the following generation are selected by " a roulette" method. And for avoiding the "failure" of Genetic Algorithm in a local minimum it is possible to add sets of standards with casual values to the sample.

Construction of one binary «genome» for application of the genetic operator is carried out on the basis of the following considerations.

One genome corresponds to one set of standards:

$$E = \{E^1, E^2, \dots, E^K\} =$$
$$= \{e_1^1, e_2^1, \dots, e_n^1; \dots; e_1^K, e_2^K, \dots, e_n^K\}.$$

As each sign has the fixed scale parameters, which are irrespective of classes, it is convenient to rearrange the components of the generalized vector of standards

$$E = \{e_1^1, e_1^2, \dots, e_1^K; \dots; e_n^1, \dots, e_n^K\}.$$

At such arrangement of components it is possible to apply identical bit sets for the group of signs for each class accordingly.

Each sign has the integral and fractional part. For calculating the number of bits of the integral part of an i-sign it is necessary to find such value $k \geq 0$, for which

$$2^{k-1} \leq [m_i] \leq 2^k, \ [m_i] \geq 1$$

$[m_i]$ is the integer part of the maximum value of a sign in training sequence.

The quantity of bits for a fractional part of a sign is calculated similarly. Value l is found, for which

$$2^{-(l+1)} \leq \{h_i\} \leq 2^{-l}.$$

$\{h_i\}$ is accuracy of approximation of i sign.

Normalized signs are applied in many systems of recognition. In such cases signs have only fractional part that simplifies the structure of Genetic Algorithm even more.

V. CONCLUSION

The Genetic Algorithm offers simple for the realization procedure of training without the teacher in the case when the quantity of classes is known and numerical signs with close stochastic characteristics are applied.

As training without the teacher can be conducted without real time restrictions (teacher must only create training sequence and indicate the quantity of classes), the Genetic Algorithm can give good approach of the optimal decision owing to multiple process of regeneration.

REFERENCES

[1] Зайченко Ю.П. "Основи проектування інтелектуальних систем. Навчальний посібник". – Київ:Видавничий Дім «Слово», 2004. – 352 с. (Zaychenko Yu. P. Basics of Intelectual Systems Designing. Tutorial. – Kyiv, 2004. – 352 p.)

[2] Фор А, "Восприятие и распознавание образов" / Пер. с фр. А.В.Серединского; под ред. Г.П.Катыса. – Москва: Машиностроение, 1989 г. – 272 с.: ил. (Faure Alain. Processing and Recognition of visual information / Tanslated from French by Seredyns'kyi A. V. – Moscow,1989. – 272 p.)

[3] Сотник С.Л., "Основы проектирования систем с искусственным интеллектом /Курс лекций". Днепродзержинск, 2000 г. http://www.codenet.ru/progr/alg/ai/htm/ (Sotnyk S. L. Basics of Artificial Intelligence Systems Designing / Course of Lectures. Dneprodzerzhinsk, 2000.)

[4] "Генетический алгоритм/ Аналитические технологии для прогнозирования и анализа данных". 2005 г. http://www.neuroproject.ru/genealg.php#begin (Genetic Algorithm/ Analytical Technologies for the Forecasting and Data Analysis, 2005) http://www.neuroproject.ru/genealg.php#begin