SELF-ORGANIZATION OF FORMAL NEURONS DESCRIBED BY THE ABSOLUTE MAXIMUM ENTROPY PRINCIPLE

I. Grabec

Faculty of Mechanical Engineering , University of Ljubljana,
P. O. Box 394, *igor.grabec@uni-lj.ac.mail.si*
61000 LJUBLJANA , Slovenia

Abstract

In this article empirical modeling of natural phenomena is treated as a mapping of sensory signals to response parameters of formal neurons. A new version of the maximum entropy principle is formulated and applied to optimization of this mapping. Its stochastic perturbation treatment leads to an adaptation process that resembles the self-organization in biological neural networks.

Introduction

The fundamental task of an intelligent system is to memorize influences from its environment and to extract information from past experiences. In biological perception and modern electronic measurement systems the influences from the environment are transformed by sensors into signals representing continuous random variables. In opposition to this the memories of biological as well as of digital electronic information processing systems are comprised of discrete elements like neurons and flip-flops. The aim of this article is therefore to answer the question: "*How can a continuous random variable be optimally mapped onto a discrete set of memory units?*" For this purpose we require that the mapping must optimally preserve the empirical information provided by sensors. This leads us in the following part of the article to the formulation of a new, so called absolute maximum entropy principle and to the derivation of an algorithm describing a self-organization process of formal neurons. (Grabec 1990)

The Maximum – Entropy Principle of Gibbs

The adaptation of a probability distribution to empirical data can be successfully performed by following the principle of maximal entropy introduced by the following problem. (Smith et al 1985) Let us consider a discrete random variable X with the sample space $S=\{x_i,\ i=1..N\}$ and let there be given some experimental data in terms of expected values of several functions g_k of random variable X

$$G_k = \sum_{i=1}^{N} p_i\, g_k(x_i) = \sum_{i=1}^{N} p_i\, g_{ki} \quad ;\ k=1,\ldots,\ K<N \tag{1}$$

The problem is to assign the probability distribution $\{p_i\}$ to the random variable X in such a way that it yields the above averages. If we have no other information about the phenomenon except the mentioned set of expected values then according to the *maximum – entropy principle of Gibbs* (Smith et al 1985) it is reasonable to select among all possible distributions that one which needs for its specification the least information. The corresponding distribution must yield the maximal entropy of information (Shannon 1948)

$$H = - \sum_{i=1}^{N} p_i \log p_i \tag{2}$$

subject to the set of constraints

$$G_k = \sum_{i=1}^{N} p_i \, g_k(x_i) = \sum_{i=1}^{N} p_i \, g_{ki} \quad ; \quad k = 1,\ldots,K<N \tag{3}$$

that also include the normalization condition.(Smith et al 1985) The maximum – entropy problem can then be solved by the calculus of variations. For this purpose we first multiply the constraints by Lagrange multipliers λ_k and form the functional

$$\mathcal{F} = - \sum_{i=1}^{N} p_i \log p_i + \sum_{k=0}^{K} \lambda_k \left(G_k - \sum_{i=1}^{N} p_i \, g_{ki}\right) \tag{4}$$

The standard variation procedure then leads to the system of equations for multipliers λ_k and probabilities $\{p_i\}$.

The absolute Maximum – Entropy Principle

Assignment of the probability distribution to a random variable on the basis of some experimental data is the central problem of empirical modeling of natural phenomena. Hence, it is not surprising that so many methods have been derived from the maximum – entropy principle, but it is surprising that only recently an essentially new version of the maximum entropy principle has emerged (Grabec 1990) which is complementary to the above formulation. In order to provide for its introduction let us discuss some properties of the solutions stemming from the procedure described above. The resulting probability distribution is generally not uniform and therefore does not correspond to the absolute maximum of the information entropy, but to a relative one. (Shannon 1948) The question therefore appears how we could assign to the phenomenon under observation a probability distribution that corresponds to the *absolute maximum* of information entropy. The above formulation of the maximum – entropy principle is logically consistent but based on the relative maximum. If we want to relay our treatment on the absolute maximum we have to change the fundamentals of the principle. We generally do not want to change the assumption about the given empirical data therefore the only possibility is to change the assumption about the fixed sample space S and to allow for an adaptable one that yields the absolute maximum of the information entropy. This means that we are looking for such a distribution of sample points as will enable assignment of uniform probability distribution to it and will still correspond to given experimental data. These assumptions differ essentially from the Gibbsian ones and we join them in *the statements of the absolute maximum – entropy principle*:

1. Let us consider a random phenomenon describable by a variable X with a continuous probability density function $f_x(x)$.

2. Let there be specified a set of reference functions $\{g(x,s)\}$, with "s" being a parameter, and let the empirical information about the phenomenon be given in terms of the mean values of these functions $\{G_e(s)=E[g(x,s)]\}$.

3. Let the sample space S_q of a representative random variable Q be comprised of K sample points $\{q_k; k=1,...,K\}$ representing disjoint elementary events, and let there be assigned to each sample point the same probability 1/K.

4. Let there exist representative mean values of reference functions $\{G_r(s)= \frac{1}{K} \sum g(q_k,s)\}$. With respect to the absolute maximum-entropy principle the discrete random variable Q optimally represents the continuous one X if some measure of discrepancy between corresponding mean values is minimal.

Here we tacitly assume the number of sample points K to be determined a priori by the design of an information processing system applied in the description of the phenomenon under observation. The positions of the sample points in the sample space $S_q=\{q_k; k=1,...,K\}$ are not fixed by the above statements but have to be placed at such positions as correspond to the third and fourth statements. The entropy of information of variable Q then corresponds to the absolute maximum. The adaptation of representative points to the probability distribution represents a problem that must be treated specifically with respect to the selected measure of discrepancy between both types of mean values. The reference functions and their mean values depend on the parameter s. When the range of this parameter is a continuous interval $(-\infty,\infty)$ we can utilize as the most simple measure for comparison of mean values the average square distance

$$D = \int [G(s) - G_r(s)]^2 \, \eta(s) \, ds \tag{5}$$

In order to ensure the convergence of this integral a weight $\eta(s)$ can generally be utilized. For the sake of simplicity we further assume $\eta(s)=1$.

The quantity D represents a measure of discrepancy between the probability distribution functions f_X and f_r pertaining to X and Q respectively. It generally depends on the position of sample points in the sample space S_q. The fundamental system of equations for the set $\{q_k ; k=1,...K\}$ is then obtained by minimizing the discrepancy D as a function of q_k :

$$\frac{\partial D}{\partial q_k} = -2 \int [G(s) - G_r(s)] \frac{\partial G_r(s)}{\partial q_k} \, ds = 0 \qquad ; k = 1,...,K \tag{6}$$

or

$$\int [G(s) - G_r(s)] \frac{\partial g(s - q_k)}{\partial q_k} \, ds = 0 \quad ; k = 1,...,K \tag{7}$$

A similar system can also be obtained for a multivariate case with s changing into a vector and ds into a corresponding differential of a volume.

Stochastic Formulation of the Absolute Maximum-Entropy Mapping

In Statement 1. we assumed that the probability density function $f_X(x)$ exists and yields the mean values of reference functions. However, in practical applications of the absolute maximum-entropy principle we have to determine

the mean values by averaging over the set of samples of the variable X in a similar way to that described in Statement 4. for the representative variable. With an increasing number of samples the estimated probability density then changes ever less, which makes feasible a perturbation treatment of the adaptation according to the absolute maximum-entropy principle. The corresponding modification of the complete treatment in a multivariate case and the related interpretation is as follows.

Let the vector $\mathbf{X}$ represent a M-component sensory signal. We treat it as a continuous random variable and describe its properties by N samples $\{\mathbf{x}_1,\ldots,\mathbf{x}_N\}$. The corresponding density of probability distribution is empirically estimated by

$$f_e(\mathbf{x}) = \frac{1}{N} \sum_{n=1}^{N} \delta(\mathbf{x} - \mathbf{x}_n) \tag{8}$$

Instead of employing an increasing set of samples we prefer to apply only a finite number of prototype vectors $\{\mathbf{q}_1,\ldots,\mathbf{q}_K\}$ for the representation of the random variable $\mathbf{X}$. A representative discrete random variable $\mathbf{Q}$ is then defined by associating a probability $1/K$ to each prototype vector : $\{P(\mathbf{q}_i)=1/K : i=1\ldots K\}$. *By this assumption the maximum entropy principle becomes the basis of the optimal mapping $X \rightarrow Q$.* If we then want to represent the probability density of the continuous variable $\mathbf{X}$ by the function

$$f_r(\mathbf{x}) = \frac{1}{K} \sum_{k=1}^{K} \delta(\mathbf{x} - \mathbf{q}_k) \tag{9}$$

we have to accommodate the prototypes to the phenomenon under observation. With this aim we compare the empirical 8 with the representative density 9 and try to diminish their discrepancy. For this purpose the singularity of the delta function is first avoided by filtering both distributions by an appropriate window function $g(\mathbf{x} - \mathbf{s})$.(Parzen 1962) Here the parameter $\mathbf{s}$ describes the window center. Filtering of both distributions then yields the empirical and the representative average:

$$\langle g \rangle_e = \int_{-\infty}^{\infty} g(\mathbf{x},\mathbf{s})\ f_e(\mathbf{x})\ d^M\mathbf{x} = \frac{1}{N} \sum_{n=1}^{N} g(\mathbf{x}_n,\mathbf{s}) \tag{10}$$

$$\langle g \rangle_r = \int_{-\infty}^{\infty} g(\mathbf{x},\mathbf{s})\ f_r(\mathbf{x})\ d^M\mathbf{x} = \frac{1}{K} \sum_{k=1}^{K} g(\mathbf{q}_k,\mathbf{s}) \tag{11}$$

The discrepancy between both distributions is described by the mean square of the difference:

$$\epsilon = \epsilon(\mathbf{q}_1,\ldots,\mathbf{q}_K ; N; \mathbf{s}) = \langle g \rangle_r - \langle g \rangle_e \tag{12}$$

$$D = \int_{-\infty}^{\infty} \epsilon^2\ d^M\mathbf{s} \tag{13}$$

The set $\{q_1,\ldots,q_K\}$ which minimizes the discrepancy D, describes an optimal mapping $X \mapsto Q$ of the continuous onto the discrete random variable.

The optimal prototypes satisfy the equations:

$$\frac{\partial D}{\partial q_{lm}} = 0 \qquad \text{or} \qquad \int_{-\infty}^{\infty} \epsilon \frac{\partial \epsilon}{\partial q_{lm}} \, d^M s = 0 \quad ; \quad l = 1...K \; ; \; m = 1 \; ... \; M \tag{14}$$

In a trivial case when N=K the optimal mapping corresponding to the absolute minimum D= 0 is determined by

$$q_1 = x_1 \; , \; q_2 = x_2 \; ... \; q_k = x_k \tag{15}$$

Otherwise the window function $g(x\text{-}s)$ must first be specified and then the system of equations 14 must be solved. The window functions $g(x\text{-}s)$ that represent a smooth approximation of delta function are generally nonlinear, therefore the system of equations 14 can not be solved in a closed form and we are forced to apply an iterative treatment. For this purpose we analyze the changes of prototypes at N >> K. With an increasing N, the empirical average $< g >_e$ converges to a fixed value, therefore we assume that at large N, the addition of successive samples causes only minor changes of prototypes. Consequently a linear approximation can be applied to express the changes of the difference ϵ which corresponds to a perturbation treatment:

$$\epsilon_1 = \epsilon(q_1 + \Delta q_1 , \ldots , q_K + \Delta q_K \; ; \; N+1 \;) = \epsilon(q_1 , \ldots , q_K \; ; \; N \;) + \; \Delta \epsilon \tag{16}$$

Where

$$\Delta \epsilon = \sum_{k=1}^{K} \sum_{i=1}^{M} \frac{\partial < g >_r}{\partial q_{ki}} \, \Delta q_{ki} + \frac{1}{N+1} \left[\, g(x_{N+1} , s) - < g >_e \right] \tag{17}$$

If the discrepancy D is minimal for $\{ q_1 , \ldots , q_k \}$ then the minimum of D_1 is obtained by the changes of prototypes $\{ \Delta q_1 , \ldots , \Delta q_k \}$ that satisfy the conditions:

$$\frac{\partial \, D_1}{\partial (\Delta q_{lm})} = 0 \quad ; \quad l = 1...K \; ; \; m = 1...M \tag{18}$$

They yield the following system of linear equations :

$$\sum_{k=1}^{K} \sum_{i=1}^{M} C_{lm\,ki} \; \Delta q_{ki} = \; B_{lm} \quad ; \; l = 1...K \; ; \; m = 1...M \tag{19}$$

The coefficients are determined by the expressions:

$$C_{lm\,ki} = \int_{-\infty}^{\infty} \frac{\partial g(q_k , s)}{\partial q_{ki}} \; \frac{\partial g(q_l , s)}{\partial q_{lm}} \; d^M s \tag{20}$$

$$B_{lm} = \frac{K}{N+1} \int_{-\infty}^{\infty} \left[\, g(x_{N+1} , s) - < g >_N \right] \frac{\partial g(q_l , s)}{\partial q_{lm}} \; d^M s \tag{21}$$

The linear system 19 can be effectively solved by iteration if the conditions

$$C_{l\,mki} = 1 \quad \text{for } k=l\ ,\ m=i \quad \text{and} \quad |\,C_{lm\,ki}\,| \ll 1 \quad \text{for } k\neq l,\ m\neq i \qquad (22)$$

could be satisfied by a proper selection of window function. In this case it is reasonable to represent the system in the form convenient for iteration :

$$\Delta q_{lm}^{(i+1)} \approx B_{lm} - \sum_{k\neq l}^{K} \sum_{i\neq m}^{M} C_{lm\,ki}\ \Delta q_{ki} \quad ;\ l = 1...K \ ;\ m = 1...M \qquad (23)$$

The iteration starts with $\Delta q_l^{(0)} = B_l$.

As a reference for the presentation of probability densities the most appropriate is the Gaussian window function :

$$g(x\text{-}s) = \exp\left[\frac{-(x-s)^2}{2\sigma^2}\right] \qquad (24)$$

With it the conditions 22 are fulfilled and the integrals in Eqs. 20 and 21 are analytically expressible as :

$$C_{lm\,ki} = \left[\ \delta_{mi} - \frac{(q_{lm}-q_{km})(q_{li}-q_{ki})}{2\sigma^2}\ \right] \exp\left[\frac{-(q_l-q_k)^2}{4\sigma^2}\right] \qquad (25)$$

$$B_{lm} = \frac{K}{N+1}\left\{(x_{N+1,m}-q_{lm})\exp\left[\frac{-(x_{N+1}-q_l)^2}{4\sigma^2}\right] - \frac{1}{K}\sum_{k=1}^{K}(q_{km}-q_{lm})\exp\left[\frac{-(q_k-q_l)^2}{4\sigma^2}\right]\right\} \qquad (26)$$

In order to fill the gap between prototypes appropriately the width of Gaussian function σ should approximately correspond to the distance between prototypes. If σ_o denotes the standard deviation of variable X, then it is reasonable to select $\sigma=\sigma_o / \sqrt{K}$. (Parzen 1962)

Numerical examples

Here we illustrate the adaptation of prototypes q_k in a one- and two-dimensional case. The generated random samples x_N are shown as thin points in Fig. 1, while the prototypes q_k are shown as thick ones. Starting from the trivial solution, the prototypes adapt to the ever less fluctuating empirical distribution of normal random variable X. The empirical and the representative cumulative probability distribution functions F of variables X and Q shown in Fig. 2 indicate the good adaptation of prototypes to the input random variable X. It is also demonstrated by the quantitative agreement between filtered probability densities $<g>_e$ and $<g>_r$ shown in Fig. 3. A slight discrepancy of both densities is mainly a consequence of the low number K. The number of iterations was three, but a similar agreement results even from a single iteration.

The adaptation of prototypes in a two-dimensional example is demonstrated in Fig. 4. The input random variable with a uniform distribution on a circle is represented by five prototypes. By marking the successive positions of prototypes during adaptation five streams of points appear that show the positioning of the prototypes. The final positions are shown in Fig. 5. Again a uniform distribution on the circle demonstrates a good agreement between the properties of the input and the representative random variable

distributions. Other examples, obtained by using various chaotic generators have shown similar performance of the derived self-organization algorithm.(Grabec 1990, 1992)

Introduction of Formal Neurons and Conclusions

The maximal entropy principle has already been applied by Gibbs (Smith et al 1985) in order to find the probability distribution function which best corresponds to a certain set of measured statistical parameters. In his approach the relative maximum of information entropy is looked for by adaptation of probabilities to given prototypes. Contrary to Gibbs we adapt the prototypes to a given probability distribution that yields the absolute maximum of information entropy. In order to emphasize this difference we call the basis of our approach *the absolute maximum-entropy principle.*

For the interpretation of the optimal maximum-entropy mapping $\mathbf{X} \mapsto \mathbf{Q}$, it is convenient to introduce the following model. Let the variable $\mathbf{X}$ represent sensory signals transmitted from the environment to the adaptive system consisting of formal neurons.(Kohonen 1988, Grabec 1990) The excitation of an individual neuron is characterized by the Gaussian function centered on the prototype $\mathbf{q}_l$. The sequence of samples $\{\mathbf{x}_N\}$ influences self-organized changes of prototypes as described by the first term of coefficient $\mathbf{B}_l$. But at the same time the neurons interact in the process of adaptation by the self-organization described by the terms depending on $\mathbf{q}_l - \mathbf{q}_k$ in expressions of $\mathbf{B}_l$ and $\mathbf{C}_{lk}$. The form of the coefficients $\mathbf{B}_l$ indicates that the input signal predominantly influences neurons having prototypes most similar to the input sample by attracting them towards this sample. The interaction between the neurons, or *self-organization*, is determined by modified bell-shaped functions consisting of excitation $(+)$ and inhibition $(-)$ parts. All these properties have been previously found in research into self-organization in some biological neural networks and artificial models resembling them. (Kohonen 1988) We can therefore conjecture that the absolute maximum entropy principle could contribute to the explanation of the properties of biological neural networks.

From the above treatment it emerges that the estimation of a continuous probability density from the discrete empirical data can be interpreted as the inverse operation to the adaptation of a discrete variable Q to a continuous probability density. From the experimental point of view the first operation corresponds to sensing and mapping of the true world to an internal picture of some information processing system while its inverse corresponds to an optimal reduction of the internal picture during its storage in a discrete memory connected to this system. The system can be either electronic or also a biological one. The corresponding procedure can be generally called *the quantization of a continuous variable.* Optimization of these operations are of importance for the development of corresponding optimal devices intended for transfer of data or communication as well as for the explanation of the reasons why the existing properties of biological neural networks have developed.

References

Grabec, I. 1990 , Biol. Cyb. **63**, 403-409

Grabec, I. 1992, in Dynamic, Genetic and Chaotic Programming, The Sixth Generation, Ed. B. Souček and IRIS Group, J. Wiley& Sons, Inc, p-p 144-163 and 470-500

Kohonen, T. 1988, Self-organization and Associative Memory,

Springer–Verlag, Berlin
Parzen, E. 1962, Ann. Math. Stat., 33,1065-1076
Shannon, C. E. 1948, Bell. Syst. Tech. J., 27 , 397
Smith, C. R. and Grandy, W. T. Jr., eds. 1985, Maximum Entropy and Bayesian Methods in Inverse Problems , D. Riedel Publ. Comp., Dordrecht

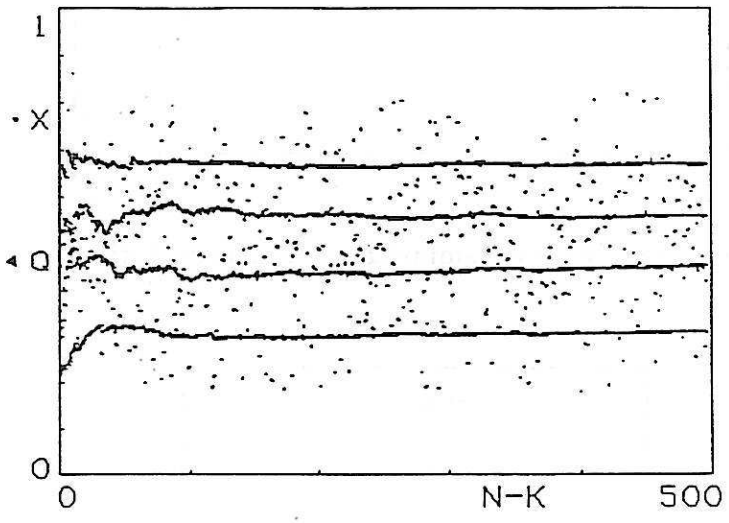

Fig. 1 The sample points (thin) and prototypes (bold) versus number of adaptations N−K . $\sigma = 0.1$

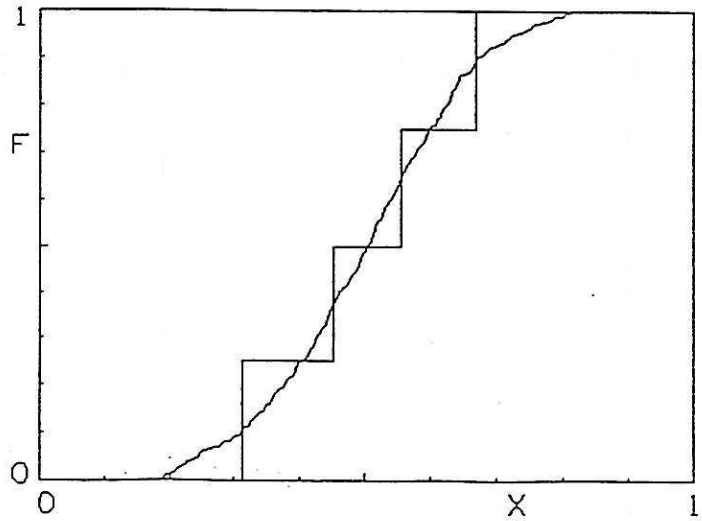

Fig. 2 Empirical (line) and representative (staircase) cumulative probability distribution functions.

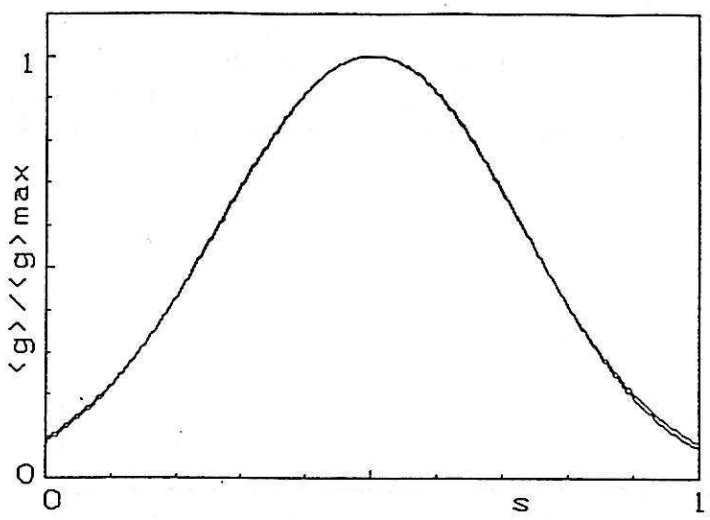

Fig. 3 The agreement between empirical and representative averages $<g>_e$ and $<g>_r$.

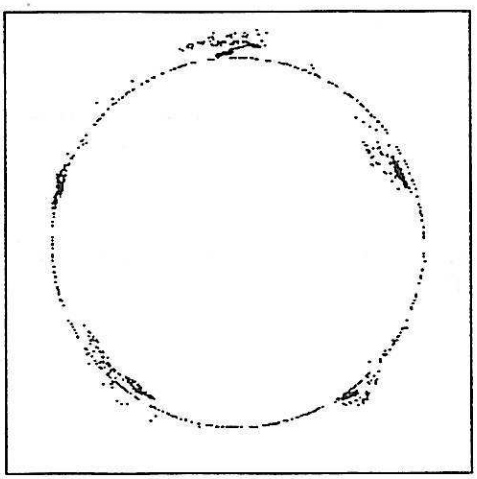

Fig.5 Adaptation of five prototypes to a circular distribution. σ=0.2

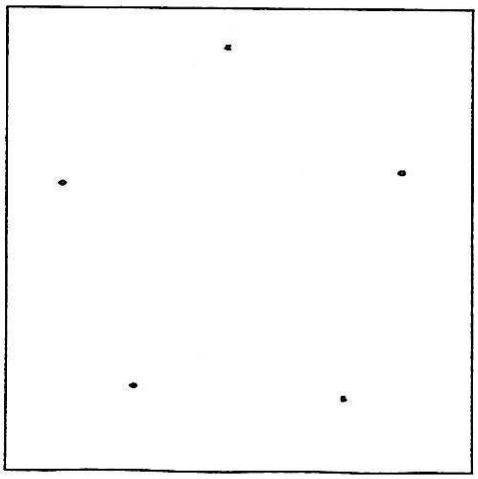

Fig.6 Final positions of five prototypes obtained after 500 adaptation steps.