INFORMATICA, 1996, Vol. 7, No. 4, 525–541

ANALYSIS OF MULTIVARIATE FUNCTION STRUCTURE IN CLASSIFICATION PROBLEMS

Vydūnas ŠALTENIS

Institute of Mathematics and Informatics
Akademijos 4, 2600 Vilnius, Lithuania
E-mail: saltenis@ktl.mii.lt

Abstract. In the present paper, the method of structure analysis for multivariate functions was applied to rational approximation in classification problems. Then the pattern recognition and generalisation ability was investigated experimentally in numerical recognition. A comparison with Hopfield Net was carried out. The overall results of using of new approach may be treated as a success.

Key words: multivariate functions, pattern recognition, neural networks, optimisation.

1. Introduction. Traditional artificial neural network (ANN) models express functions $f(x_1, \ldots, x_n)$ of several variables as a composition of basic functions, called units or network nodes. The units are parametrized as a nonlinear transformation of linear combination of many variables. The attractiveness of the network methods is that functions of high dimensions, after the optimisation of parameters, are often closely approximated by compositions of lower dimensional functions.

There are three ways of thinking about neural networks. The *descriptive* line of thought is concerned with the proximity of the artificial model to biological systems. The *computational* approach views this model as a novel computational paradigm. Finally, there is the *normative* view of neural networks, which examines the mathematical and statistical backdrop of neural architectures and learning algorithms. The last view attempts to analyse where neural networks either violate or extend what can be done by other, traditional models.

The normative approach is nearest to the paper. Indeed, any brain performs much better than any computer when recognising real objects. But it is doubtful if natural principles and algorithms are always efficient because of a deep discrepancy between biological systems and modern computers.

It is common knowledge, that ANN, trained by the back propagation algorithm, require a great number of sweeps of the training sample data in order to minimise the classification error. A long training time is one of the principal characteristics of the back propagation classifiers. ANN training is very slow, because optimisation of weights is a highly dimensional and multiextremal problem (Vyšniauskas, 1994).

It has been proven mathematically that almost any classification problem can be simulated by a neural network with only one hidden layer of neurones (Hornik, 1991). But the mathematical proof doesn't tell us how many neurones the hidden layer should include. The construction of effective topology of ANN is still more art than science.

The aim of this paper is to avoid these two imperfections. We apply the analysis of multivariate functions, select the main groups of variables and use them in the approximation by functions of fewer variables for classification problems.

There is great difficulty in solving multidimensional problems (multiextremal optimisation, evaluation of integrals) in the case of high dimensionality. However, the investigation of practical problems shows that the influence of variables or their groups is extremely different. Frequently we may suggest a separation of a relatively small part of the main variables or their groups. The efficiency of such simplification algorithms depends on the structure of the problem.

2. Decomposition into components of different dimensionality. A decomposition of a multivariate function into the summands of different dimensionality (Cukier *et al.*, 1979; Šaltenis, 1989; Sobolj, 1990) makes the base for the structure analysis.

Let a function

$$f(X) = f(x_1, \dots, x_n),$$

be defined, for simplicity, on the unit cube K^{n} $(0 \leqslant x_1 \leqslant 1, \dots, 0 \leqslant 1)$:

$$X \in K^n.$$

Sometimes a short notation f will be used for $f(x_1, \dots, x_n)$.

Let us introduce notation for the domain set of the function f, which is a Cartesian product of basic domains $\Omega_1, \dots, \Omega_n$:

$$\Omega = \Omega_1 \times \cdots \times \Omega_n,$$

and for special domains:

$$\begin{aligned}
&\Omega_{i_1 \dots i_s} = \Omega_{i_1} \times \cdots \times \Omega_{i_s}, \quad 1 \leqslant i_1 < \dots < i_s \leqslant n, \ s = 1, \dots, n,\\
&\Omega_{(i)} = \Omega_1 \times \cdots \times \Omega_{i-1} \times \Omega_{i+1} \times \cdots \times \Omega_n, \quad i = 1, \cdots, n,\\
&\Omega_{(ij)} = \Omega_1 \times \cdots \times \Omega_{i-1} \times \Omega_{i+1} \times \cdots \times \Omega_{j-1} \times \Omega_{j+1} \times \cdots \times \Omega_n,\\
&\quad i, j = 1, \dots, n, \quad i < j.
\end{aligned}$$

In the general case, the domain $\Omega_{(i_1 \dots i_s)}$ is defined in a similar way.

We shall use groups of indices $i_1, \dots, i_s$, where $1 \leqslant i_1 < \dots < i_s \leqslant n$, $s = 1, \dots, n$ and denote the sum with $2^n - 1$ terms as:

$$\sum^{\wedge} T_{i_1 \dots i_s} = \sum_{i=1}^{n} T_i + \sum_{1 \leqslant i < j \leqslant n} \sum T_{ij} + \cdots + T_{12 \dots n}.$$

The decomposition of the function f into summands of different dimensionality

$$f = f_0 + \sum^{\wedge} f_{i_1 \dots i_s}(x_{i_1}, \dots, x_{i_s}) \tag{1}$$

is unique and orthogonal for each function f integrable on K^n (Sobolj, 1990), if f_0 is constant and the integrals of summands (1) are equal to zero:

$$\int_0^1 f_{i_1 \dots i_s}(x_{i_1}, \dots, x_{i_s}) dx_{i_k}, \quad 1 \leqslant k \leqslant s. \tag{2}$$

The summands of decomposition (1) may be found just like some integrals.

Let us introduce the following notation for the function of s variables:

$$f^{i_1 \dots i_s} = \int_{\Omega(i_1 \dots i_s)} f,$$

where the superscripts $i_1, \dots, i_s$ of f indicate, that the integral is over all basic domains except $i_1, \dots, i_s$.

Then, after integrating of (1) on Ω, the constant sumand will be equal to

$$f_0 = \int_\Omega f, \tag{3}$$

one-dimensional summands, after integrating on $\Omega_{(i)}$, will be equal to

$$f_i(x_i) = f^i - f_0, \quad i = 1, \dots, n,$$

two-dimensional summands, after integrating on $\Omega_{(ij)}$, will be equal to

$$f_{ij}(x_i, x_j) = f^{ij} - f_0 - f_i(x_i) - f_j(x_j), \quad i, j = 1, \dots, n, \quad i < j$$

and so on.

3. Approximation by functions of fewer variables. We approximate a multivariate function f by the approximating function f_e, which is the sum of unknown functions of fewer variables, and seek to minimise the approximation error

$$\|f - f_e\| = \left(\int_\Omega (f - f_e)(f - f_e) \right)^{1/2}.$$

The unknowns, in the case, are not the coefficients but the functions and they are found from functional equations (Golomb, 1959; Šaltenis, 1989).

f_e may consist of functions of various dimensionality. It is interesting that the functions in decomposition (1) may be used as the best approximations in the case.

The approximating function of zero order $f_e^{(0)}$ is a constant f_0.

The approximating function of first order $f_e^{(1)}$ is a sum of one-dimensional functions:

$$f_e^{(1)} = \sum_{i=1}^{n} f_i + f_e^{(0)}$$

In the general case

$$f_e^{(s)} = \sum_{1 \leqslant i_1 < \dots < i_s \leqslant n} f_{i_1 \dots i_s} + \sum_{r=0}^{s-1} f_e^{(r)}, \quad s = 1, \dots, n-1.$$

4. Structure characteristics and approximation errors. The system of structure characteristics:

$$D = \overset{\wedge}{\sum} D_{i_1 \ldots i_s},$$

$$D_{i_1 \ldots i_s} = \int\limits_{\Omega} \left(f_{i_1 \ldots i_s}\right)^2, \tag{4}$$

$$D = \int\limits_{\Omega} (f)^2 - (f_0)^2 \tag{5}$$

was proposed, investigated (Šaltenis, 1989) and applied in analysing the structure of optimisation problems. The structure characteristics $D_{i_1 \ldots i_s}$ indicate the degree of influence of the respective variable groups in approximation.

Let us include into the approximation those summands of fewer variables, which groups of indices $\{i_1, \ldots, i_s\}$ make up the set I_e. Then the approximation error is equal to

$$\delta = \|f - f_e\|^2 = \sum_{i_1 \ldots i_s \notin I_e} D_{i_1 \ldots i_s} \tag{6}$$

For example, if the function of two variables $f(x_1, x_2)$ is approximated by the function of one variable $f_e(x_1)$, the approximation error δ will be equal to $D_2 + D_{12}$.

Relationship (6) is useful when choosing the variables and their groups for rational approximation by means of selecting some number N_D of the greatest structure characteristics. This enables us to approximate multivariate functions by functions of fewer variables with the minimal error.

We must also keep in mind that in real situations evaluations of structure characteristics are usually with significant errors so only the greatest characteristics may be more accurate.

The abbreviation MUSTAN (MUltivariate STructure ANalysis) will be used for this computational model.

5. Evaluation of structural characteristics. If we know the values of the function $f(X)$ for some points X^j $(j = 1, \ldots, N)$, then the Monte-Carlo method is used for the evaluations basing on (3), (4) and (5):

$$f_0 \approx \frac{1}{N} \sum_{j=1}^{N} f(X^j), \tag{7}$$

$$D+(f_0)^2 \approx \frac{1}{N}\sum_{j=1}^{N}\left(f(X^j)\right)^2, \tag{8}$$

where N is the number of samples,

$X^j = (x_1^j, \ldots, x_n^j)$ are random points of dimensionality n, uniformly distributed in Ω.

s coordinates $Y^j = (x_{i_1}, \ldots, x_{i_s})$ must be equal for pairs of random points used for the evaluation of $D_{i_1 \ldots i_s}$:

$$D_{i_1 \ldots i_s} + \left(f_0\right)^2 \approx \frac{1}{N}\sum_{j=1}^{N} f(Y^j, Z^j) f(Y^j, U^j),$$

where Y^j are random points of dimensionality s, uniformly distributed in $\Omega_{i_1 \ldots i_s}$,

Z^j and U^j are random points of dimensionality $n-s$, uniformly distributed in $\Omega_{(i_1 \ldots i_s)}$.

6. The case of two level values of variables. The values of variables x_i may be treated, for example, as values of brightness of some pattern points. Each $X = (x_1, \ldots, x_n)$ must be classified as belonging to one of K patterns: $P^1, \ldots, P^K$.

It is natural to introduce such a function for each pattern:

$$g^k(X) = \begin{cases} 1 & \text{if } X \text{ must be identified as } P^k, \; k = 1, \ldots, K, \\ 0 & \text{otherwise.} \end{cases}$$

If we have some approximations g_e^k for functions g^k, then the classification of the given pattern values X may be carried out according to the condition:

$$\max_{k=1,\ldots,K} g_e^k(X). \tag{9}$$

Let the values of variables x_i be equal to 0 or 1. Then each basic domain consists of two points $\Omega_i = \{0,1\}$, all domain Ω consists of 2^n points. So the integrals in our equations must be changed into the respective sums.

Let us introduce a function for two level argument values:

$$\chi(x) = \begin{cases} -1 & \text{if } x = 0, \\ 1 & \text{if } x = 1. \end{cases}$$

Then decomposition (1) may be changed into:

$$f = f_0 + \overset{\wedge}{\sum} c_{i_1 \ldots i_s} \chi(x_{i_1}) \ldots \chi(x_{i_s}), \tag{10}$$

where $c_{i_1 \ldots i_s}$ are coefficients.

The decomposition (10) is unique and orthogonal because the summands

$$f_{i_1 \ldots i_s} = c_{i_1 \ldots i_s} \chi(x_{i_1}) \ldots \chi(x_{i_s})$$

satisfy the condition (2). Really, the mean value of $\chi(x)$ for each basic domain is equal to zero so the sums respective to integrals of condition (2) are also equal to zero.

Then the domain $\Omega_{i_1 \ldots i_s}$ may be divided in two parts:

$$\Omega_{i_1 \ldots i_s} = \Omega^+_{i_1 \ldots i_s} + \Omega^-_{i_1 \ldots i_s}$$

according to the sign of $\chi(x_{i_1}) \ldots \chi(x_{i_s})$:

$$\begin{aligned} X &\in \Omega^+_{i_1 \ldots i_s} \quad \text{if} \quad \chi(x_{i_1}) \ldots \chi(x_{i_s}) = 1, \\ X &\in \Omega^-_{i_1 \ldots i_s} \quad \text{if} \quad \chi(x_{i_1}) \ldots \chi(x_{i_s}) = -1. \end{aligned}$$

The mean values of f on domains $\Omega^+_{i_1 \ldots i_s}$ and $\Omega^-_{i_1 \ldots i_s}$ are equal respectively to:

$$\begin{aligned} \frac{1}{2^{n-1}} \sum_{X \in \Omega^+_{i_1 \ldots i_s}} &= f_0 + c_{i_1 \ldots i_s}, \\ \frac{1}{2^{n-1}} \sum_{X \in \Omega^-_{i_1 \ldots i_s}} &= f_0 - c_{i_1 \ldots i_s}, \end{aligned}$$

because the mean values of summands, different from $f_{i_1 \ldots i_s}$ are equal to zero.

Then

$$c_{i_1 \ldots i_s} = \frac{1}{2^{n-2}} \left(\sum_{X \in \Omega^+_{i_1 \ldots i_s}} f - \sum_{X \in \Omega^-_{i_1 \ldots i_s}} f \right). \tag{11}$$

It is clear that:

$$D_{i_1 \ldots i_s} = (c_{i_1 \ldots i_s})^2, \quad 1 \leqslant i_1 < \ldots < i_s \leqslant n, \quad s = 1, \ldots, n. \tag{12}$$

Let us introduce notation for the Monte-Carlo evaluations of respective characteristics: $\overline{f}_0, \overline{D}, \overline{c}_{i_1\ldots i_s}, \overline{D}_{i_1\ldots i_s}$ of the function g^k for pattern k in the case of two-level variables. They may be evaluated after using (7), (8), (11) and (12) in such a way:

$$\overline{f}_0 \approx \frac{1}{N}\sum_{j=1}^{N} g^k(X^j) = \frac{N^k}{N},$$

$$\overline{D} = \overline{f}_0(1-\overline{f}_0),$$

$$\overline{c}_{i_1\ldots i_s} = \frac{1}{2}\left(\frac{1}{N^+_{i_1\ldots i_s}} \sum_{X^j\in\Omega^+_{i_1\ldots i_s}} g^k(X^j) - \frac{1}{N^-_{i_1\ldots i_s}} \sum_{X^j\in\Omega^+_{i_1\ldots i_s}} g^k(X^j)\right),$$

$$\overline{D}_{i_1\ldots i_s} = \left(\overline{c}_{i_1\ldots i_s}\right)^2,$$

where N, X^j are the same as for (7), (8) in chapter 5,

N^k is the number of points that must be classified as belonging to pattern k,

$N^+_{i_1\ldots i_s}$, and $N^-_{i_1\ldots i_s}$ are the numbers of summands in the respective sums.

7. Phases of calculation

- Evaluations of characteristics $\overline{f}_0, \overline{D}, \overline{c}_{i_1\ldots i_s}, \overline{D}_{i_1\ldots i_s}$ for all patterns $k = 1, \ldots, K$, if we know the values of the function $g^k(X)$ for some points X^j $(j = 1, \ldots, N)$. Our main assumption is that the points of the samples used in the "learning" phase are really uniformly and independently distributed in domain Ω.
- A number N_D of the greatest characteristics are selected and approximations g^k_e for functions g^k are constructed:

$$g^k_e = \overline{f}_0 + \sum_{1}^{N_D} \overline{c}_{i_1\ldots i_s}\chi(x_{i_1})\cdots\chi(x_{i_s}).$$

  Usually, the structure characteristics $D_{i_1\ldots i_s}$ of only limited order s $(s < 4)$ are evaluated.
- The "recall" phase consists in calculating the values of g^k_e for new points X and selecting the largest approximating function values according to condition (9).

8. A pattern recognition example. Experiments and results. In order to estimate the abilities of MUSTAN, an experiment was carried out on the recognition of numerals because most successful ANN applications have been reported in recognition problems (for example, Fukushima, 1988; Widrow and Winter, 1988).

Digitised character data was a 10 × 8 pixel bi-level image. The i-th pixel corresponds to the variable x_i.

Usually the experiments may be classified into two broad categories (Hertz *et al.*, 1991): *recognition* and *generalisation problems.* In the *recognition* problem, pairs of input and output $(I_1, O_1),\ (I_2, O_2), \ldots, (I_m, O_m)$ were used for training the network and the trained network was tested by the input $I_1, I_2, \ldots, I_m$, corrupted by noise. The network is expected to reproduce the output O_j corresponding to I_j in spite of the presence of noise.

In *generalisation* problems, the network was tested by input I_{m+1} , which was distinct from the inputs $I_1, I_2, \ldots, I_m$ used for training. The network is expected to correctly predict the output O_{m+1} for the previously unseen input I_{m+1}.

Some samples (without noise and corrupted by noise), used in recognition experiments are shown in Fig. 1. Noise level is a probability of changing the original noncorrupted value for each point of a pattern.

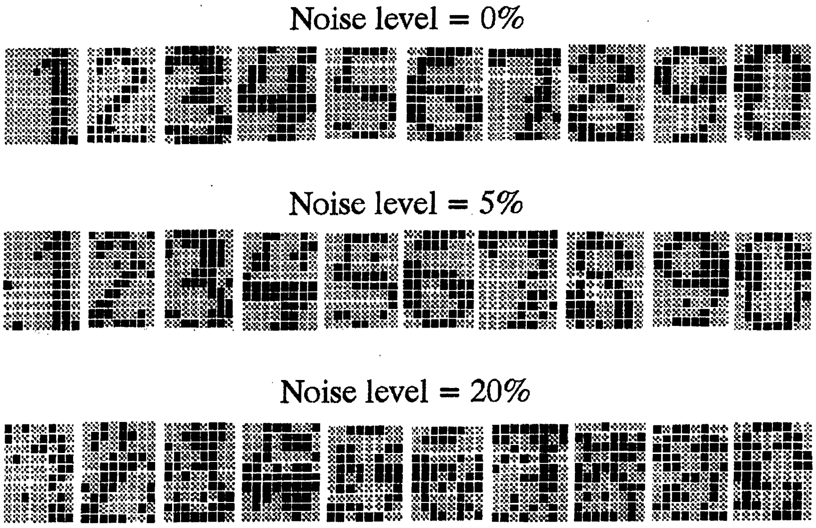

Fig. 1. Samples of numerals used for recognition experiments.

Some samples of numeral "2" used in generalisation experiments are shown in Fig. 2.

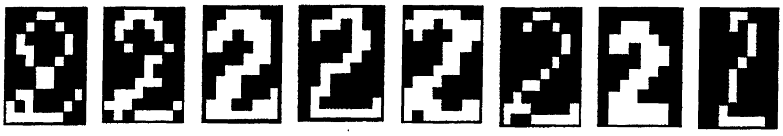

Fig. 2. Samples of numerals used for generalisation experiments.

Using MUSTAN in recognition experiments, the author examined the recognition rate in cases where the number of main structural characteristics N_D was equal to 2, 5, 10, 30, 80, 150, 3240, and for noise levels equal to 0, 5, 20, and 40%. The results of recognition experiments are shown in Fig. 3.

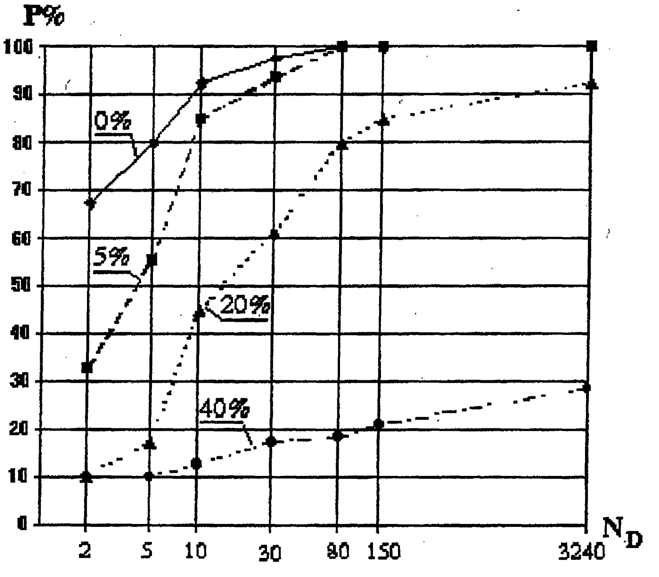

Fig. 3. Recognition rate by number of structural characteristics N_D for different noise levels.

The results of generalisation experiments are shown in Fig. 4.

The number of experiments for evaluating of recognition rate was equal to 60–450 depending on situation.

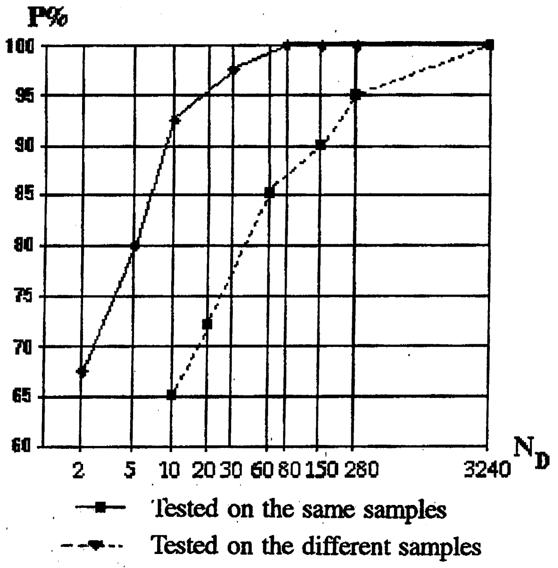

Fig. 4. Recognition rate by number of structural characteristics N_D tested on the same samples (recognition case) and on different samples (generalisation case).

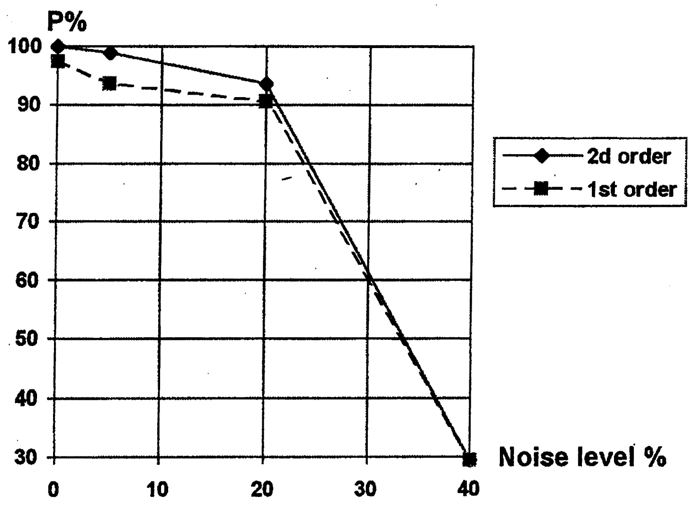

Fig. 5. Recognition rate by noise level when structure analysis was limited by first order and second order characteristics.

The influence of limitation of structure analysis by some maximal order s is investigated by special experiments. Their results are shown in Fig. 5.

The following remarks can be made from the experiments:

- The recognition rate achieved using a small part of structural characteristics is relatively high.
- Only a high noise level (over 20%) can considerably influence the recognition reliability.
- Fig. 4 shows the generalisation ability of MUSTAN.

9. Some interpretations of the results. It is of interest to have some visualised results of structure analysis of approximation functions g_e^k. Some coefficients c_i of the greatest single structure characteristics for the approximation function of numeral "2" are visualised in Fig. 6.

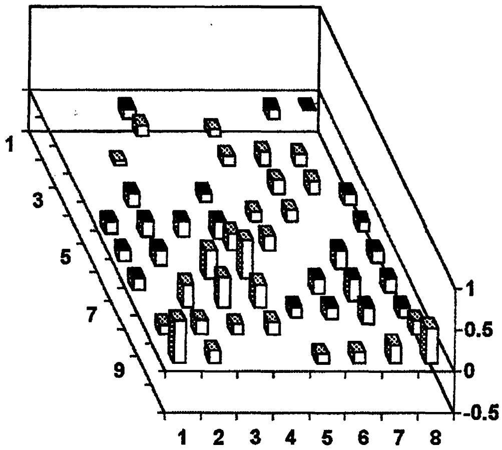

Fig. 6. Graphical representation of the greatest single structure characteristics for the approximation function of numeral "2".

We may see that the greatest positive coefficients are located in the most typical to the numeral "2" points (bottom-left side of "2"). The negative coefficients are on the right–middle height places, where the points of "2" occur rarely.

The greatest double structure characteristics may be interpreted as some links between two variables. Therefore they were visualised in Fig. 7 in the form of lines, connecting the respective points. Solid lines are used for positive coefficients c_{ij} , dash lines – for negative ones. Links from the upper right to the bottom left corner are really typical of numeral "2".

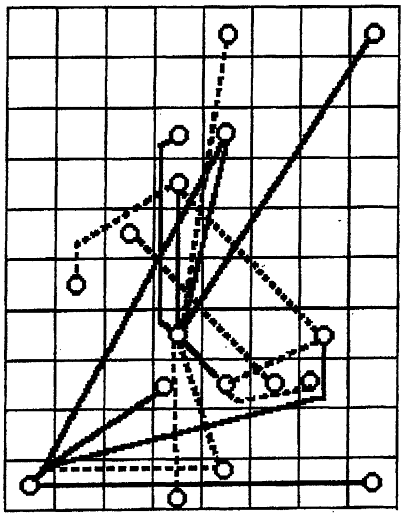

Fig. 7. Graphical representation of the greatest double structure characteristics for the approximation function of numeral "2".

Fig. 8 is an attempt to interpret the greatest structural characteristics of first, second and third order as some nodes, similar to ANN.

10. Experimental comparison with the Hopfield Net. Cross-bar Associative Network (Hopfield, 1982), which is usually referred to as a Hopfield Net, was used to solve the same numeral recognition problems. The number of processing elements in the Hopfield Net was equal to 80 (the number of pixels in the image). The Net was able to successfully learn only four numeral images, therefore experiments with MUSTAN were conducted with the same four numerals.

Fig. 9 illustrates the recognition rates for various noise levels.

Generalisation experiments, similar to that of Chapter 7, were carried out in order to compare the Hopfield Net and MUSTAN abilities. The results are presented in Table 1.

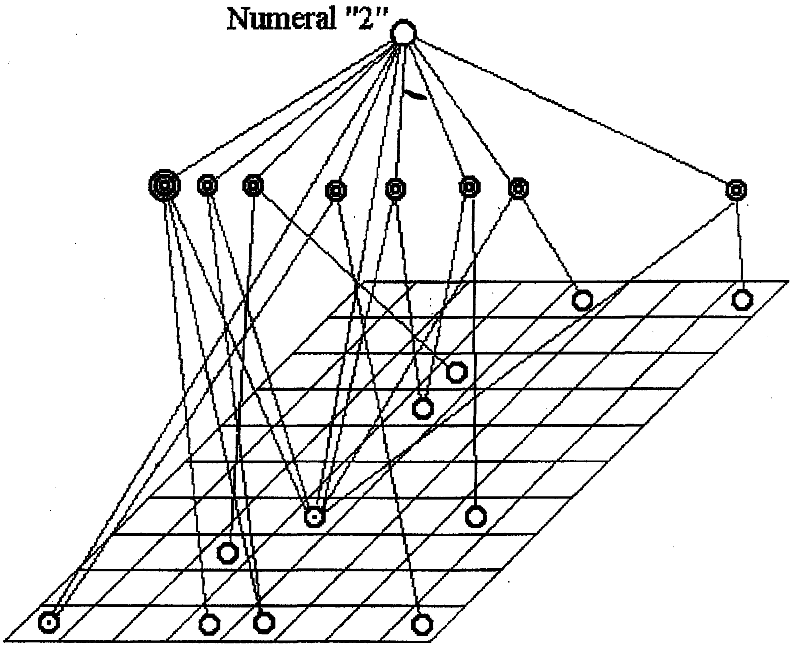

Fig. 8. The nodes, equivalent to the structural characteristics of: first order ⊙ , second order ◎ , third order ◉

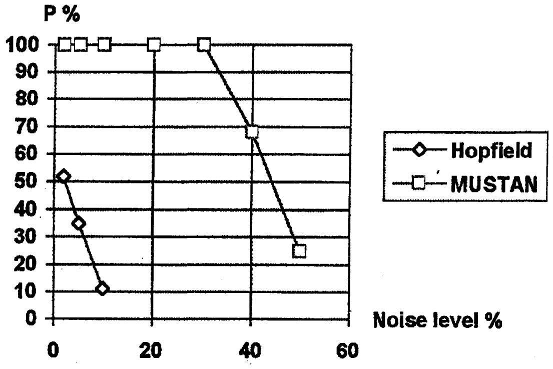

Fig. 9. Comparison of the recognition rate with the Hopfield Net.

Table 1. Comparison of recognition rates for Hopfield Net and MUSTAN in the generalisation experiments

Recognition algorithms	Recognition rate %
Hopfield MUSTAN	16.3 76.7

MUSTAN outperforms the Hopfield Net both in recognition and generalisation.

11. Discussion. We may see some correspondence between the MUSTAN approach and ANN. The main structural groups of variables in MUSTAN, in a sense, look like ANN basic nodes after training. The difference is that we do not involve the laborious optimisation of weights but use a relatively simple evaluation of numerical parameters in MUSTAN. We also have no problems with the construction of an effective topology of ANN.

Obviously, the evaluation of *all* structural characteristics of high order is practically impossible for relatively high dimensionality n. For example, if $n = 100$, the number of structural characteristics of third order is equal to 161700 and the number of structural characteristics of fourth order is equal to 3921225. This drawback may be surmounted by using some properties of structural characteristics.

First of all, it is known that structure characteristics of relatively high order are negligibly small in real problems.

Another property is also promising. Some statistical links exist between the values of high and low order. After the analysis of first order characteristics we may efficiently decrease the search for the greatest double characteristics and so on.

12. Conclusions. The main point of this paper was to show that the approach of structure analysis offers a promise in some cases traditional to ANN. The overall results of new MUSTAN procedure may be treated as successful. Future studies will determine the extent to which the promise of the approach is actually fulfilled.

13. Acknowledgements. The author wishes to thank his colleagues V. Vyšniauskas and V. Tiešis for helpful discussions.

REFERENCES

Cukier, R.I., H.B. Levine and K.E. Shuler (1979). Nonlinear sensitivity analysis of multiparameter model systems. *Journal of Computational Physics* , **26**(1), 1–42.

Fukushima, K. (1988). A neural network for visual pattern recognition. *IEEE Computer,* March, 65–75.

Golomb, M. (1959). Approximation by functions of fewer variables. In R.E. Langer (Ed.), *On Numerical Approximation.* The University of Wisconsin Press, Madison. pp. 275–327 .

Hertz, J., A. Krogh and R.G. Palmer (1991). *Introduction in the Theory of Neural Computing.* Addison Wesley.

Hopfield, J. (1982). Neural Networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences,,* **79**, 2554–2558.

Hornik, K. (1991). Approximation capabilities of multilayer feedforward networks. *Neural Networks,* **4**, 251–257.

Šaltenis, V. (1989). *Structure Analysis of Optimisation Problems.* Mokslas publishers, Vilnius (in Russian).

Sobolj, I.B. (1990). On sensitivity estimation for nonlinear mathematical models. *Matematitcheskoye Modielirovanye,* **2**(1), 112–118 (in Russian).

Vyšniauskas, V. (1994). Searching for minimum in neural networks. *Informatica,* **5**(1–2), 241–255.

Widrow, B., and R. Winter (1988). Neural nets for adaptive filtering and adaptive pattern recognition. *IEEE Computer,* March, 25–39.

Received November 1996

V. Šaltenis graduated from the Kaunas Technological Institute, Lithuania, in 1959. He received Ph.D. degree from the Moscow Energy Institute of the USSR Academy of Sciences in 1966. He is a senior researcher of the Optimization Department at the Institute of Mathematics and Informatics, Lithuania. Present research interests include both theory and applications of the structure of optimisation problems, multicriteria decision support systems.

DAUGELIO KINTAMŲJŲ FUNKCIJŲ STRUKTŪROS ANALIZĖ KLASIFIKAVIMO UŽDAVINIUOSE

Vydūnas ŠALTENIS

Klasifikavime, taikant dirbtinius neuroninius tinklus, iškyla komplikuoti optimizavimo uždaviniai, parenkant didelio kintamųjų skaičiaus funkcijų parametrus – vadinamuosius svorius. Darbe siūloma analizuoti šių funkcijų struktūrą, išskiriant labiausiai įtakojančias nedidelio kintamųjų skaičiaus grupes, kurių vertinimas būtų palyginti paprastas. Atlikti eksperimentiniai tyrimai, taikant struktūros analizę atpažįstant skaitmenų vaizdus. Tirta atsitiktiniais triukšmais iškraipytų skaitmenų, o taip pat skaitmenų, nenaudotų mokymo pavyzdžiuose, atpažinimo galimybės. Atliktas eksperimentinis palyginimas su Hopfield'o dirbtiniu neuroniniu tinklu.