

# Evolutionary Method with Clustering for Feature Selection

S.A Subbotin, An.A. Olejnik

*Faculty of software, Zaporozhye National Technical University,  
Zhukovskiy Street, 64, Zaporozhye, Ukraine*

subbotin@zntu.edu.ua, olejnik@zntu.edu.ua

**Abstract.** *Application of evolutionary search for the decision of a feature selection problem is considered. The evolutionary method with feature clustering, raising efficiency of search of the most significant feature combination due to use of the aprioristic information of a feature arrangement in the space of copies is offered. Results of feature selection for the decision of a problem of model construction of aircraft engines hardening factor are shown.*

## Key words

Evolutionary search, feature selection, neural network, significant feature.

## 1 Introduction

The maintenance of high reliability of functioning, effective and safety work of complex technical objects causes the necessity of the automated systems development for diagnostics and non destroying quality assurance, that demands the decision of a problem of the recognizing models construction describing controllable objects or processes. However real technical objects, as a rule, are characterized by a plenty of controlled variables, some of which are superfluous and not informative, that complicates the process of diagnostic model construction, results in its redundancy and downturn interpretability, that further increases time of forecasting or classification on synthesized model [1].

Therefore the important stage at the model construction of multivariate objects and processes is a dimension reduction of researched feature space and a choice of the most informative feature system [2]. Nowadays various methods of feature selection [3] are known. Among them are: feature ranging, exhaustive search, heuristic methods, methods of evolutionary search.

At the solving of forecasting and recognition problems, as a rule, it is necessary to deal with a system of statistically dependent features, which significance is not expressed through significance of separate features [4]. Therefore at the solving of problems of diagnostics it is necessary to estimate a feature set, instead of each feature separately. Thus, the usage of feature ranging by the calculated individual estimation is unacceptable.

The usage of an exhaustive search related with the necessity of estimation of all possible combinations of the features that makes impossible application of such approach at a plenty of features in an initial set as demands huge computing expenses [5].

Methods of heuristic search [6] are insufficiently effective because of nonoptimality of the greedy strategy that consistently adding or excluding one feature therefore the obtained feature set contains the superfluous features correlating with other features in a set. Besides that at the feature selection from a set with high dimension, heuristic search also demands significant expenses for an estimation of feature sets.

For a selection of the most informative feature combination in conditions of their interdependence it is represented expedient to choose evolutionary search [7–9] as it is more adapted to a finding of new solutions due to association of the best decisions obtained on various iterations, has opportunities for an output from local optimum, uses group estimations of selfdescriptiveness of a feature set instead of individual.

However classical methods of evolutionary search at the feature selection do not take into account affinity of an arrangement of features in a space of copies therefore the new combinations of features (chromosome) formed by application of evolutionary operators of initialization, crossover and mutation, can include features which contain the identical information on researched object, process or system. It is obvious, that the feature sets corresponding to such chromosomes, are low informative or superfluous.

The purpose of this work is a creation of an evolutionary method with feature clustering which takes into account an arrangement of features in a space of copies at creation of new solutions and allows to allocate a combination of the informative features belonging to different factorial groups.

## 2 Evolutionary method with feature clustering

In the developed method of evolutionary search with feature clustering it is offered to group similar features with the help of methods of clustering which allow to share sample into groups of compactly located features in space of copies (clusters, factorial groups) and to allocate in every cluster the most typical feature.

At the creation of new chromosomes using evolutionary operators of initialization, crossover and mutation it is offered to calculate the probability of inclusion of a feature into a chromosome, which depends on a feature arrangement in a cluster (distances from it to the center of cluster), individual significance of a feature, and also individual significance of its cluster center.

Evolutionary search with feature clustering is offered to be performed as the following sequence of steps.

Step 1. Group features of initial sample in clusters.

Step 1.1. For each feature  $X_i$  calculate Euclidean distance from it to all other features in sample. Euclidean distance  $d_E(X_a; X_b)$  between features  $X_a$  and  $X_b$  is calculated using the formula:

$$d_E(X_a; X_b) = \sqrt{\sum_{p=1}^m (x_{pa} - x_{pb})^2},$$

where  $x_{pa}$  and  $x_{pb}$  are values of  $a$ -th and  $b$ -th features of  $p$ -th copy of training sample, respectively.

Step 1.2. Generate groups of the features compactly located in space of copies based on calculated distances between copies, using methods of the cluster-analysis [10, 11]. Allocate features being the centers of clusters.

Step 1.3. For each feature  $X_i$  calculate the probability of its inclusion into a chromosome.

Step 1.3.1. Calculate the value of individual significance  $I_i$  feature  $X_i$ .

Step 1.3.2. To define {determine} probability  $P_i$  of inclusion of  $i$ -th feature into a chromosome:

$$P_i = I_i + \frac{d_E(X_i; X_{c,i})}{d_{E_{\max,c}}}(I_i - I_c),$$

where  $d_E(X_i; X_{c,i})$  is the distance from feature  $X_i$  to its center of cluster;  $d_{E_{\max,c}}$  is the maximal distance in cluster in which  $i$ -th feature is located;  $I_c$  is the significance of a feature being the center of cluster in which feature  $X_i$  is located.

Step 2. Set the counter of iterations:  $t = 0$ .

Step 3. Initialize an initial population from  $N$  chromosomes. Thus the chromosome is represented by bit line of size  $L$ , where  $L$  is the quantity of features in an initial set. If the gene of a chromosome accepts value "1" then the feature corresponding to it is considered informative and it is taken into account at the estimation of a feature set corresponding to a chromosome. Otherwise, when the gene of a chromosome accepts zero value, the feature is considered non informative and is not used at the estimation of a combination of features.

Step 3.1. Set the counter of the generated chromosomes:  $j = 1$ .

Step 3.2. Generate  $j$ -th chromosome  $H_j$ .

Step 3.2.1. Set the counter of the certain genes:  $i = 1$ .

Step 3.2.2. Generate a random number:  $r = \text{rand}[0; 1]$ .

Step 3.2.3. If  $P_i > r$  then for  $i$ -th gene of  $j$ -th chromosome to set value:  $h_{ij} = 1$ , otherwise:  $h_{ij} = 0$ .

Step 3.2.4. If  $j$ -th chromosome is generated completely ( $i = L$ ) then go to a step 3.3.

Step 3.2.5. Set:  $i = i + 1$ .

Step 3.2.6. Go to a step 3.2.2.

Step 3.3. If all chromosomes are generated ( $j = N$ ) then go to a step 4.

Step 3.4. Set:  $j = j + 1$ .

Step 3.5. Go to a step 3.2.

Step 4. Calculate the value of fitness-function  $f'(H_j)$  of chromosomes of the current population using the formula:

$$f'(H_j) = \frac{f(H_j) \sum_{i=1}^L h_{ij}}{\left(1 + \sum_{i=1}^L I_i h_{ij}\right) \left(1 + \sum_{i=1}^L P_i h_{ij}\right)}.$$

Step 5. Check the termination criteria. If criteria of the termination are satisfied then go to a step 11.

Step 6. Increase the counter of iterations:  $t = t + 1$ .

Step 7. Select chromosomes for crossover and mutation.

Step 8. Execute the operator of uniform crossover. Thus in a crossover mask set individual values for genes to which there correspond features with probability of inclusion in a chromosome, is higher than average, to other genes to appropriate zero values.

Step 9. Execute the operator of a point mutation. The probability of mutation  $P_{Mi}$  of  $i$ -th gene in a mutated chromosome is offered to calculate using the formula:  $P_{Mi} = \alpha (1 - P_i)$ , where  $\alpha$  is the factor determining degree of a mutation,  $\alpha \in [0; 1]$ .

Step 10. Generate new generation. Go to a step 4.

Step 11. Stop.

In the offered method of evolutionary search with feature clustering it is taken into account the affinity of an arrangement of features in a space of copies that allows to form new decisions of the features located, as a rule, in different groups, increasing probability of search of a combination of the features with maximal significance.

### 3 Experiments and results

For the research of properties and characteristics of the offered method the problem of feature selection for model construction of hardening factor of aircraft engine details was solved.

As the factors describing the process of diamond hardening of details in [12] it is offered to use 16 features determining geometrical, physical and technical characteristics of a processable detail and process of hardening.

Feature selection was performed based on methods of evolutionary search. Thus the following evolutionary operators were applied: the selection operator – roulette selection, the crossover operator – uniform crossover, the mutation operator – point mutation. Parameters of search were established by the following: the quantity of individuals in population  $N = 100$ , the quantity of elite individuals is three, the probability of crossover  $P_c = 0,8$ . Termination criteria: maximal allowable quantity of iterations  $T = 100$ , achievement of comprehensible value of criterion function  $f_a = 0,01$ .

At the feature selection as a fitness-function  $f$  for estimation of a feature set (chromosome) the expression was used:

$$f(H_j) = \left(1 + \frac{1}{L} \sum_{i=1}^L h_{ij}\right) I_j,$$

in which the significance of feature combination  $I_j$  is calculated as the sum squared error of the model constructed based on selected features. As such model it was used feed forward three-layer neural network which first layer contained six neurons, second layer contained three neurons and the third layer contained one neuron, all neurons of the network had sigmoid activation function.

Results of the experiments are shown in tab. 1, where  $t$  is the time spent for evolutionary search of an informative feature combination, sec.;  $k_f$  is the quantity of the calculated values of fitness-function;  $k$  is the quantity of the selected features.

**Tabl.1.** Results of feature selection.

Method of feature selection	Criterion			
	$t$	$k_f$	$k$	$f$
Classical evolutionary search	821,72	3155	12	0,0152
Evolutionary search with кластеризацией features	717,82	2837,9	12	0,0093

Results of experiments have shown, that at the usage of the offered method of evolutionary search with feature clustering for selection of an informative combination of features is spent less time and it is required less calculations of fitness-function in comparison with classical evolutionary search. Thus the optimum set contains less features and allows to synthesize neural network model, providing best accuracy.

## 4 Conclusions

In this work the actual problem of an informative feature selection is solved based on evolutionary approach.

Scientific novelty of the is that the new method of evolutionary search with feature clustering in which the affinity of an arrangement of features in a space of copies is taken into account is developed. It allows to generate new solutions from the features located, as a rule, in different groups, increasing probability of search of a combination of the features with the greatest significance.

Practical value of results of the work is that the problem of feature selection for model synthesis of dependence of hardening factor of aircraft engine details is solved.

## References

- [1] Елисеева И.И., Рукавишников В.О. Группировка, корреляция, распознавание образов (Статистические методы классификации и измерения связей). – М.: Статистика, 1977. – 143 с.
- [2] Биргер И.А. Техническая диагностика. – М.: Машиностроение, 1978. – 240 с.
- [3] Интеллектуальные средства диагностики и прогнозирования надежности авиадвигателей: Монография / В.И. Дубровин, С.А. Субботин, А.В. Богуслаев, В.К. Яценко. – Запорожье: ОАО "Мотор-Сич", 2003. – 279 с.
- [4] Айвазян С.А., Енюков И.С., Мешалкин Л.Д. Прикладная статистика: Исследование зависимостей. – М.: Финансы и статистика, 1985. – 487 с.
- [5] Guyon I., Elisseeff A. An Introduction to Variable and Feature Selection // Journal of Machine Learning Research. – 2003. – №3. – P. 1157–1182.
- [6] Dash M., Liu H. Feature Selection for Classification // Intelligent Data Analysis. – 1997. – №1. – P. 131–156.
- [7] Букатова И.Л., Михасев Ю.И., Шаров А.М. Эвоинформатика: Теория и практика эволюционного моделирования. – М.: Наука, 1991. – 206 с.
- [8] Эволюционные методы компьютерного моделирования: Монография / А.Ф. Верлань, В.Д. Дмитриенко, Н.И. Корсунов, В.А. Шорох. – К: Наукова думка, 1992. – 256 с.
- [9] Рутковская Д., Пилиньский М., Рутковский Л. Нейронные сети, генетические алгоритмы и нечеткие системы: Пер. с польск. И.Д. Рудинского. – М.: Горячая линия – Телеком, 2004. – 452 с.
- [10] Классификация и кластер / Под ред. Дж. Вэн Райзина. – М.: Мир, 1980. – 392 с.
- [11] Жамбю М. Иерархический кластер-анализ и соответствия. – М.: Финансы и статистика, 1988. – 342 с.
- [12] Богуслаев В.А., Яценко В.К., Притченко В.Ф. Технологическое обеспечение и прогнозирование несущей способности деталей ГТД. – К.: Манускрипт, 1993. – 333 с.