UDC 004.93

**S. SUBBOTIN[1], A. OLIINYK[1], V. LEVASHENKO[2], E. ZAITSEVA[2]**

[1] *Zaporizhzhya National Technical University, Ukraine*
[2] *University of Zilina, Faculty of Management Sciences and Informatics, Slovakia*

# INDUCTION OF CLASSIFICATION RULES IN CASE OF UNEVEN DISTRIBUTION OF CLASSES

*The problem of induction of classification rules on the basis of negative selection in the case of uneven distribution of classes in the sample is solved. The new method for the induction of such rules is proposed. This method uses a priori information about instances of all classes in the sample. A hypercube of maximum possible volume is used as a form of detector. It allows to exclude irrelevant and redundant features from the sample, thereby reducing the search space and the time of execution of the method, as well as to generate a set of detectors with high approximation and generalization capability. The software implementing the proposed method is developed. Some experiments on the solution of problem of gas turbine air-engine blade diagnosis are conducted.*

***Key words:*** *artificial immune system, classification rules, misclassification error.*

## Introduction

Process of building decision models for pattern recognition is actual task [1, 2]. Situation when the most data of training set belong to one class is a typical for such process [2]. We have to develop new models for object's formalization or process description. One of the perspective approaches for develop such models based on conception of artificial immune systems [3-5]. This model can be created based on one class only. The difference between numbers of instances, which belonging to different classes is significant in this case. Usage of artificial immune systems with negative selection is proposed in [6-8]. These systems involve the construction of a set of detectors that are capable of recognizing unknown instances [9-10]. This approach allows to detect anomalies or random variations in diagnosed objects [3, 6], and to recognize instances of "non-self" classes (classes of objects which are not represented in the training set) [4, 7, 9]. There are known methods for the synthesis of artificial immune systems based on the negative selection [4-10]. These methods generate an exhaustive number of detectors (possible solutions) and employ instances with one class only. Instances with other classes do not taken into account.

Consequently, the development of methods for the synthesis of artificial immune systems, which are free from these disadvantages, is an actual task. Diagnostic models based on artificial immune systems have a low level of generalization. The detectors (rules) of the immune system are easy in understanding. However, because of the low level of generalization, a detector system has a large dimension. It is difficult to understand and analyze by human, which generally leads to reduction of interoperability of the diagnostic model.

So, the purpose of this paper is developing a method of induction of classification rules based on the basis of a set of detectors. These rules handles data of training set, with a significant difference in the number of instances which belong to different classes.

## 1. Problem statement

Let us assume that there is a training set S=<P, T>, where P is a set of input parameters (features) of an objects and set T is a set of values of the output parameter. Set of values of the output parameter is represented as a vector $T = (t_q)_Q$, where $t_q \in T'$ is a value of the output parameter of q-th instance; T' is a set of possible values of the output parameter (usually in problems of non-destructive quality control and pattern recognition a set T' consists of two elements $T' = \{t'_0, t'_1\}$, determining class of suitability of object, such if $t_q = t'_0$ then q-th object is considered unusable, if $t_q = t'_1$ it is usable, suitable etc.).

The number of instances of the sample one class (for example, instances of the class $t_q = t'_1$) is significantly different from the number of instances of another class, which is defined as

$$0 \le N_{t_q = t'_0} << N_{t_q = t'_1}, \text{ where } N_{t_q = t'_0} + N_{t_q = t'_1} = Q.$$

Then, on the basis of a training set S=<P, T> it is necessary to generate a set $\{rule_1, \ldots, rule_{NR},\}$ of

productions $P_r \rightarrow T_r$, that allows to provide an acceptable level of misclassification error E. This error E defined as the ratio of number of incorrectly recognized instances $N_{er}$ to the total number of instances Q.

## 2. The method of classification rule

Known methods of negative selection [4-10] have such disadvantages as the generation of the exhaustive number of detectors, the usage of information of one class instances only, low interoperability of synthesized set of solutions of detectors etc. To eliminate these disadvantages it is advisable to use the method of classification rules synthesis on the basis of negative selection in the case of uneven distribution of instances of the sample classes. In this method is used:

– a known information about instances of both classes $T' = \{t'_0, t'_1\}$ in generating the set of detectors $AB = \{Ab_1, Ab_2, \ldots, Ab_{Nab}\}$. It is forming a set of detectors with high approximation and generalizing properties;

– an information about individual significance of features $p_m$. It is eliminating irrelevant and redundant features of the sample S=<P, T>;

– a hypercube of maximum possible volume as a form of detector. It is a contrast to known methods of negative selection, in which a hypersphere is used as a form of detector. This hypercube allows to eliminate the necessity of solving a resource intensive problem of search of optimal radius of hyperspheres of detectors.

Evaluation of the significance of features $p_m$ with respect of the output parameter $T$ is the initial stage of proposed method. It allows to identify and to exclude irrelevant features from further consideration, thereby reducing the search space and time of the method.

The proposed method for the classification rules synthesis based on negative selection approach. This method is oriented to the case of uneven distribution of class instances of sample in generating a set of detectors. The proposed method uses known information about instances of all classes of the sample. It also takes into account information about the individual significance of features [11-13]. A hypercube of maximum possible volume is used as a form of detector.

It allows to exclude irrelevant and redundant features from the sample, thereby reducing the search space and time of the method implementation. As result, a set of detectors with high approximation and generalization capability is formed.

So, the proposed method increases the generalizing properties of synthesized model by reducing the number of detectors and conditions of antecedents. This method improves interoperability of model, reduces its dimension (structural and parametric complexity) and volume of used memory. All of these improving are increasing model performance with sequential computation.

## 3. Experiments and results

Software has been developed for implementation of proposed method. This software is oriented to the analysis of different characteristics of the method. Its deals with a problem of blade diagnosis for gas turbine of aircraft engine [14]. Blades of gas turbine were characterized by the values of the power spectra of damped oscillations after impact excitation. These values of the power spectra are used as input features. Classes of blade quality were defined with the help of experts: undamaged and defective (potentially dangerous). Each blade was described by 10240 characteristics of the power spectrum of damped oscillations. Artificial features were constructed to reduce the search space based on these characteristics. A set consisting of 80 artificial features was obtained based on this reducing.

The resulting sample S=<P, T> does not have statistical representativeness, because it does not display the actual frequency distribution of classes. Really, number of undamaged blades is substantially greater in the general population than the number of defective blades. These defective blades ($t_q = t'_1$) in the sample represent typical cases of nonconformity, which provides a topological representation of defective blades in the sample. All the possible cases of the class of undamaged blades ($t_q = t'_0$) cannot be present in a sample from a practical point of view. Therefore, it is necessary to build a diagnostic model for aircraft engine blade class recognition based on the available sample S=<P, T> with uneven distribution of classes' instances.

The sample S=<P, T> contains 42 instances characterizing defective blades and 72 instances representing undamaged blades. The proposed method for the synthesis of classification rules was compared with the existing methods of negative selection. These methods synthesized a set of detectors based on "self" instances $S_1 \subseteq S$ of the sample only. The problem of blade diagnosis of gas turbine of aircraft engines was solved with the proposed method two times:

– using a sub-sample $S_1 \subseteq S$, which contains information about defective instances ("self") only;

– using all original samples S=<P, T>.

Experimental investigation of characteristics proposed method has been compared with other methods of negative selection.

The experimental results are given in Table. 1. This table contains next values. Column $N_{it}$ describes a number of iterations of the method. Misclassification

error on the training data S=<P, T> indicated at column E. Column $E_t$ contains misclassification error of the test data. Columns $P_{t,1-0}$ and $P_{t,0-1}$ indicated probability of misclassification. So, value $P_{t,1-0}$ is error probability of assignment to class "self" ($t_q$=t'$_1$) when instance actually belongs to a class of "non-self" ($t_q$=t'$_0$). Similarly, value $P_{t,0-1}$ is error probability of assignment to class "non-self" ($t_q$=t'$_0$) when an instance actually belongs to a class of "self" ($t_q$=t'$_1$).

Table 1

Experimental results

| Method | $N_{Ab}$ | $N_{it}$ | E | $E_t$ | $P_{t,1-0}$ | $P_{t,0-1}$ |
|---|---|---|---|---|---|---|
| RNS [8] | 207 | 50 | 0.070 | 0.136 | 0.126 | 0.333 |
| MVD [9] | 41 | 50 | 0.035 | 0.077 | 0.069 | 0.250 |
| MMD [10] | 19 | 14 | 0.018 | 0.055 | 0.054 | 0.083 |
| MPRSBNS ($S_1 \subseteq S$) | 20 | 12 | 0.026 | 0.037 | 0.038 | 0.000 |
| MPRSBNS (S=<P, T>) | 31 | 19 | 0.009 | 0.011 | 0.011 | 0.000 |

Table 1 shown that the misclassification error E values produced by the method MMD [10] (E=0.018) and by the proposed method MPRSBNS (E = 0.026 and E = 0.009) are acceptable. The low recognition errors of these methods were provided by wide coverage of field of "self" instances $S_1 \subseteq S$ by synthesized detectors. The proposed method MPRSBNS has synthesized a set of detectors based on instances of all classes of the sample S=<P, T>. This method has provided more acceptable results (E = 0.009) compared a set of detectors synthesized using "self" instances $S_1 \subseteq S$ (E = 0.026) only. Method RNS [8] and model V-Detector [9] had less acceptable misclassification error (E = 0.070 and E = 0.035, respectively). This fact indicates the lack of coverage by the synthesized detectors the area of "self" instances $S_1 \subseteq S$. The experimental results show that the method RNS [8] and the model V-Detector [9] generate the largest number of detectors ($N_{Ab}$ = 207 and $N_{Ab}$ = 41, respectively). Method MMD [10] and the proposed method MPRSBNS (using sample $S_1 \subseteq S$) have generated significantly fewer number of detectors ($N_{Ab}$ = 19 and $N_{Ab}$ = 20, respectively). It indicates a more efficient operation of these methods. In particular, the method MPRSBNS uses a priori information about the significance of features at the initial stage. This method eliminates from further consideration irrelevant and redundant features that can reduce the search space and create a set of a small number of detectors based on highly informative features with a high approximation and generalization capability.

Criteria $E_t$, $P_{t,1-0}$ and $P_{t,0-1}$ were used for the analyzing of investigated methods. These criteria describe misclassification error and the probability of making a wrong decision based on test data. Misclassification error of models synthesized by the proposed method MPRSBNS and methods [8-10] are shown in Table 1. These errors have been calculated on test data $E_t$. Misclassification error of the proposed method MPRSBNS is significantly lower than the error of other known methods ($E_t$ = 0.136, $E_t$ = 0.077 and $E_t$ = 0.055 for the methods [8-10], respectively). The method MPRSBNS allowed to reach misclassification error $E_t$ = 0.037 (using a part of the sample $S_1 \subseteq S$) and $E_t$ = 0.011 (using the full sample S=<P, T>).

It is important to note the specificity of the solved problem of blade diagnosis. An error of assignment to "non-self" class ($t_q$ = t'$_0$) has a very high cost when instance actually belongs to a "self" class ($t_q$ = t'$_1$). This error evaluates by criterion $P_{t,0-1}$. This is due to the fact that the classification of defective blades to the class of undamaged can cost human lives. The test data has zero error probability $P_{t,0-1}$ for the proposed method MPRSBNS. This fact indicates high efficiency of the proposed method for solving such problems. The zero level of error probability $P_{t,0-1}$ by using the proposed method is explained by a high level of coverage of typical instances of class $t_q$ = t'$_1$. This coverage was made by generated set of detectors AB ={$Ab_1$,…,$Ab_{Nab}$}. Note, that this set of detectors was obtained with using a priori information about the importance of features;

Thus, the results of experiments showed that the proposed method due to the usage of a priori information and exclusion of irrelevant and redundant features of the sample makes it possible to reduce the search space and time of execution. Proposed method allows to synthesize classification models in a form of a set of detectors with high approximation and generalization capabilities. Also by reducing the number of detectors and the conditions in antecedents it increases interpretability of the model, reduces its dimension and, therefore, the size of the used memory.

## Conclusion

In this paper we solve the urgent problem of automation of classification rule synthesis based on negative selection for the case of uneven class distribution in the sample.

The developed method of classification rule synthesis based on negative selection uses a priori information about instances of all classes in the sample at detector set generation. It also takes into account information about the individual feature significance. An experimental study of the proposed method and its comparison with the known analogues is performed. A practical task of diagnosing the vanes of gas turbine of aircraft engines has been solved. The mathematical

approach proposed at [15] can be used for reliability analysis of the proposed solution.

## References (GOST 7.1:2006)


*1. Sobhani-Tehrani, E. Fault diagnosis of nonlinear systems using a hybrid approach / E. Sobhani-Tehrani, K. Khorasani. – New York : Springer, 2009. – 265 p. – (Lecture notes in control and information sciences ; № 383).*

*2. Gertler, J. Fault detection and diagnosis in engineering systems / J. Gertler. – New York : Marcel Dekker Inc., 1998. – 483 p.*

*3. Mo, H. Handbook of research on artificial immune systems and natural computing: applying complex adaptive technologies / H. Mo. – London : Information Science Reference, 2009. – 634 p.*

*4. Tarakanov, A. O. Immunocomputing: principles and applications / A. O. Tarakanov, V. A. Skormin, S. P. Sokolova. – New York : Springer, 2003. – 230 p.*

*5. Dasgupta, D. Immunological computation: theory and applications / D. Dasgupta, F. Nino. – Boca Raton : Auerbach Publications, 2008. – 296 p.*

*6. Ji, Z. Revisiting negative selection algorithms / Z. Ji, D. Dasgupta // Evolutionary Computation. – 2007. – Vol. 15. – P. 223-251.*

*7. Zhang, P. T. A malware detection model based on a negative selection algorithm with penalty factor / P. T. Zhang, W. Wang, Y. Tan // Science China. Information sciences. – 2010. – Vol. 53. № 12. – P. 2461-2471.*

*8. Oliveira, L. O. Real-valued negative selection (RNS) for classification task / L. O. Oliveira, I. N. Drummond // Recognizing patterns in signals, speech, images and videos. – 2010. – Vol. 6388. – P. 66-74.*

*9. Ji, Z. Negative selection algorithms: from the thymus to V-detector : thesis ... doctor of philosophy in "Computer Science" / Zhou Ji. – Memphis, 2006. – 337 p.*

*10. Zaitsev, S. A. Negative selection using masked detectors / S. A. Zaitsev, S. A. Subbotin // Моделирование неравновесных систем : XIV Всероссийский семинар, Красноярск, 7-9 октября 2011 г.: материалы. – Красноярск : СФУ, 2011. – С. 95-98.*

*11. Oliinyk, A. Training Sample Reduction Based on Association Rules for Neuro-Fuzzy Networks Synthesis / A. Oliinyk, T. Zaiko, S. Subbotin // Optical Memory and Neural Networks (Information Optics). – 2014. – Vol. 23, № 2. – P. 89–95.*

*12. Oliinyk, A. O. Using Parallel Random Search to Train Fuzzy Neural Networks / A. O. Oliinyk, S. Yu. Skrupsky, S. A. Subbotin // Automatic Control and Computer Sciences. – 2014. – Vol. 48, Issue 6. – P. 313–323.*

*13. Oliinyk, A. O. Experimental Investigation with Analyzing the Training Method Complexity of Neuro-Fuzzy Networks Based on Parallel Random Search / A. O. Oliinyk, S. Yu. Skrupsky, S. A. Subbotin // Automatic Control and Computer Sciences. – 2015. – Vol. 49, Issue 1. – P. 11–20.*

*14. Интеллектуальные средства диагностики и прогнозирования надежности авиадвигателей : моногр. / В. И. Дубровин, С. А. Субботин, А. В. Богуслаев, В. К. Яценко. – Запорожье : ОАО "Мотор-Сич", 2003. – 279 с.*

*15. Kvassay, M. Evaluation of Algorithms for Identification of Minimal Cut Vectors and Minimal Path Vectors in Multi-State Systems / M. Kvassay, J. Kostolny // Communications - Scientific Letters of the University of Žilina. – 2015. – Vol. 174. – P. 8-14.*


## References (BSI)


1. Sobhani-Tehrani, E., Khorasani, K. *Fault Diagnosis of Nonlinear Systems using a Hybrid Approach*. New York, Springer Publ., 2009. 265 p.

2. Gertler, J. *Fault Detection and Diagnosis in Engineering Systems.* New York, Marcel Dekker Inc. Publ., 1998. 483 p.

3. Mo, C. L. *Handbook of Research on Artificial Immune Systems and Natural Computing: Applying Complex Adaptive Technologies*. London, Information Science Reference Publ., 2009. 634 p.

4. Tarakanov, A., Skormin, V., Sokolova, S. *Immunocomputing: Principles and Applications*. New York, Springer Publ., 2013. 230 p.

5. Dasgupta D. *Immunological Computation: Theory and Applications*. Boca Raton, Auerbach Publ., 2008. 296 p.

6. Ji, Z., Dasgupta, D. Revisiting Negative Selection Algorithms. *Evolutionary Computation*, 2007, no. 15, pp. 223-251.

7. Zhang, P.T., Wang, W., Tan, Y. A Malware Detection Model Based on a Negative Selection Algorithm with Penalty Factor. *Science China. Information sciences*, 2010, vol. 53, no.12, pp. 2461-2471.

8. Oliveira, L., Drummond, I. Real-valued Negative Selection (RNS) for Classification Rask. *Recognizing Patterns in Signals, Speech, Images and Videos*, 2010, no. 6388, pp. 66-74.

9. Ji, Z. *Negative Selection Algorithms: From the Thymus to V-detector.* Doctor of philosophy in "Computer Science", thesis, Memphis, 2006. 337 p.

10. Zaitsev, S., Subbotin, S. Negative Selection using Masked Detectors. *Proc. of seminar on Modelling of non-equilibrium systems*, Krasnoyarsk, 2011, pp. 95-98.

11. Oliinyk, A., Zaiko, T., Subbotin, S. Training Sample Reduction Based on Association Rules for Neuro-Fuzzy Networks Synthesis. *Optical Memory and Neural Networks,* 2014, vol. 23, no. 2, pp. 89-95.

12. Oliinyk, A., Skrupsky, S., Subbotin, S. Using Parallel Random Search to Train Fuzzy Neural Networks. *Automatic Control and Computer Sciences*, 2014, vol. 48, no. 6, pp. 313-323.

13. Oliinyk, A., Skrupsky, S., Subbotin, S. Experimental Investigation with Analyzing the Training

Method Complexity of Neuro-Fuzzy Networks Based on Parallel Random Search. *Automatic Control and Computer Sciences*, 2015, vol. 49, no. 1, pp. 11-20.

14. Dubrovin, V., Subbotin, S., Boguslayev, A., Yatsenko, V. *Yntellektual'nye sredstva dyahnostyky y prohnozyrovanyya nadezhnosty avyadvyhateley* [Intelligent Diagnosis and Prediction of Reliability of Aircraft Engines]. Zaporozhye, Motor-Sich Publ., 2003. 279 p.

15. Kvassay, M., Kostolny, J. Evaluation of Algorithms for Identification of Minimal Cut Vectors and Minimal Path Vectors in Multi-State Systems, *Communications - Scientific Letters of the University of Žilina*, 2015, vol. 174, pp. 8-14.

*Поступила в редакцию 15.03.2016, рассмотрена на редколлегии 14.04.2016*

## СОЗДАНИЕ ПРАВИЛ КЛАССИФИКАЦИИ ПРИ НЕРАВНОМЕРНОМ ЗАДАНИИ КЛАССОВ

***С. А. Субботин, А. А. Олейник, В. Г. Левашенко, Е. Н. Зайцева***

В работе рассматривается проблема построения правил классификации на основе отрицательного отбора в случае неравномерного распределения классов в исходной выборке. Авторами предлагается новый метод построения таких правил. Этот метод использует априорную информацию о распределении классов в выборке, в качестве формы детектора использует гиперкуб максимально возможного объема, что позволяет исключать малозначимые и избыточные признаки из выборки, сократив тем самым пространство поиска и время выполнения метода, а также формировать набор детекторов с высокими аппроксимационными и обобщающими способностями. Разработано программное обеспечение для реализации предложенного метода. Представлены экспериментальные результаты диагностики лопаток газотурбинного авиадвигателя.

**Ключевые слова:** искусственная иммунная система; правила классификации; ошибка классификации.

## ПОБУДОВА ПРАВИЛ КЛАСИФІКАЦІЇ ПРИ НЕРІВНОМІРНОМУ ЗАВДАННІ КЛАСІВ

***С. О. Субботін, А. О. Олійник, В. Г. Левашенко, О. М. Зайцева***

У роботі розглядається проблема побудови правил класифікації на основі негативного відбору для випадку нерівномірного розподілу класів у вихідній вибірці. Авторами пропонується новий метод побудови таких правил. Цей метод при генерації набору детекторів використовує відому інформацію про екземпляри всіх класів вибірки, враховує інформацію про індивідуальну значущість ознак, як форму детектора використовує гіперкуб максимально можливого об'єму, що дозволяє виключати малозначущі та надлишкові ознаки з вибірки, скоротивши тим самим простір пошуку і час виконання методу, а також формувати набір детекторів з високими апроксимаційними та узагальнюючими здібностями. Розроблено програмне забезпечення для реалізації запропонованого методу. Наведено експериментальні результати діагностування лопатей газотурбінного авіадвигуна.

**Ключові слова:** штучна імунна система; правила класифікації; помилка класифікації.

**Субботин Сергей Александрович** – д-р техн. наук, проф., заведующий кафедрой программных средств, Запорожский национальный технический университет, Запорожье, Украина, e-mail: subbotin.csit@gmail.com.

**Олейник Андрей Александрович** – канд. техн. наук, доцент, доцент кафедры программных средств, Запорожский национальный технический университет, Запорожье, Украина, e-mail: olejnikaa@gmail.com.

**Левашенко Виталий Григорьевич** – канд. физ.-мат. наук, проф., Университет г.Жилина, Словакия, e-mail: vitaly.levashenko@fri.uniza.sk.

**Зайцева Елена Николаевна** – канд. физ.-мат. наук, проф., Университет г.Жилина, Словакия, e-mail: elena.zaitseva@fri.uniza.sk.

**Sergey Subbotin** – Dr. Sc., Prof., Head of Software tools department, Zaporizhzhya National Technical University, e-mail: subbotin.csit@gmail.com

**Andrii Oliinyk** – PhD, Assoc. Prof., Assoc. Prof. of Software tools department, Zaporizhzhya National Technical University, e-mail: olejnikaa@gmail.com

**Vitaly Levashenko** – PhD, Prof., University of Zilina, Slovakia, e-mail: vitaly.levashenko@fri.uniza.sk.

**Elena Zaitseva** – PhD, Prof., University of Zilina, Slovakia, e-mail: vitaly.levashenko@fri.uniza.sk.