ITM Web of Conferences **6**, 02001 (2016)
DOI: 10.1051/itmconf/20160602001
© Owned by the authors, published by EDP Sciences, 2016

Using of the notion «Pareto set» for development of the forecasting models based on the modified clonal selection algorithm

Nadezhda Astakhova[1] and Liliya Demidova[1, 2, a]

[1] *Ryazan State Radio Engineering University, 390005 Ryazan, Russia*
[2] *Moscow Technological Institute, 119334 Moscow, Russia*

Abstract. The algorithm which carries out the multiobjective optimization at realization of the modified clonal selection algorithm based on the use of the notion «Pareto set» when the parental population of antibodies should be created for development of the forecasting models on the base of the strictly binary trees has been offered. Two indicators of quality of the forecasting model – the affinity indicator based on the calculation of the average forecasting error rate, and the tendencies discrepancy indicator – are applied in the role of the objective functions. The results of experimental studies which confirm the efficiency of application of the offered algorithm have been given.

1 Introduction

The key stage at the solution of the forecasting problem of time series (TS) with use of the artificial intelligence technologies is the choice's stage of the best forecasting model. In particular, in the forecasting model based on the strict binary trees (SBT) and the modified clonal selection algorithm (MCSA) [1, 2] the forecasting model is presented in the form of antibodies. The antibody is a sequence encoded by randomly selected characters. Such sequence will be transformed to the analytical dependence which represents some function. This function will be applied to obtaining the predicted values of TS.

The search of the best forecasting model occurs during process of iterative calculations. The best forecasting models must be determined at each step of this process. Such models become parents for the next generation of models at the next step [1 – 8]. Obviously, the correct selection of antibodies is the key for the effective use of the MCSA and its convergence.

The traditional approach to choose the short-term forecasting models of TS consists in the quality estimation of the forecasting models by means of the average forecasting error rate, calculated for the training set of data. Herewith the average forecasting error rate should be minimized [1 – 8].

However, the use of the average forecasting error rate as the single quality indicator of the forecasting model is not always sufficient to determine the best forecasting model. Often it is required to consider the additional quality indicators of the forecasting model, such as the compliance to the seasonal tendencies of TS, the compliance to the trend of TS, lack of emissions, complexity of the forecasting model, etc. [3].

[a] Corresponding author: liliya.demidova@rambler.ru

Usually it isn't possible to choose the single quality indicator. Therefore, the development problem of approaches to creation of the forecasting models conforming to the requirements about providing an extremum of several quality indicators is very actual.
Therefore, it is expedient to use the additional quality indicator which will allow to estimate the general tendency of values' change of the known elements of TS (for example, the tendencies discrepancy indicator) along with the average forecasting error rate [3]. Hence, it is possible to increase the efficiency of the forecasting models on the base of the SBT at the solution of the problem of medium-term forecasting.

2 Theoretical part

The average forecasting error rate $AFER$ [1, 2], which is also called the affinity indicator Aff (in the context of working with the MSCA) and used as one of the quality indicators for the forecasting models can be calculated as:

$$AFER = \frac{1}{n-r}\sum_{j=r+1}^{n} |(f^j - d^j)/d^j| \cdot 100\%, \quad (1)$$

where d^j and f^j are respectively the actual (fact) and forecasted values for the j-th element of TS (for the j-th timing); n is the number of TS elements (number of timing).

The rate of discrepancy between the tendencies of TSs (the tendencies discrepancy indicator $Tendency$) is used

This is an Open Access article distributed under the terms of the Creative Commons Attribution License 4.0, which permits unrestricted use distribution, and reproduction in any medium, provided the original work is properly cited.

as other quality indicator for the forecasting models and can be calculated as:

$$Tendency = \frac{h}{n-r-1}, \quad (2)$$

where h is the number of negative multiplications $(f^{j-1} - f^j) \cdot (d^{j-1} - d^j)$; $j = \overline{r+2,n}$; d^j and f^j are respectively the actual (fact) and forecasted values for the j-th element of TS (for the j-th timing); n is the number of TS elements (number of timing); r is the model order; $n-r-1$ is the total number of multiplications $(f^{j-1} - f^j) \cdot (d^{j-1} - d^j)$.

This indicator is used for adaption of the forecasting models on the base of the SBT and MCSA for the medium-term forecasting.

Both indicators (Aff and $Tendency$) determine the similarity of the predicted values of the analyzed TS with the real ones. However, they use different principles of evaluation. The affinity indicator Aff is used in the implementation of the MCSA to define value of «adaptability» (quality) of the antibody Ab, and the tendencies discrepancy indicator $Tendency$ allows to estimate the quality of the antibody Ab taking into account the coincidence with the trend of the analyzed TS. Herewith both indicators must be minimized.

These indicators are based on various principles of the quality assessment of the forecasting model. The affinity indicator Aff estimates the similarity and difference between the predicted and actual values of the known elements of the analyzed TS. The tendencies discrepancy indicator $Tendency$ estimates the similarity and difference of the change's directions between the predicted and actual values of the known elements of the analyzed TS. This indicator helps to analyze tendencies in the TS and the presence of seasonal fluctuations.

Thus, both indicators (the affinity indicator (1) and the tendencies discrepancy indicator (2)) must be used simultaneously at the quality assessment of the forecasting models on the base of the SBT and MCSA to solve the tasks of medium-term forecasting.

Various well proved approaches can be applied to the solution of the task of the simultaneous accounting of two quality indicators for development of the forecasting models [9]. Herewith it is necessary especially to allocate approach, based on the several multiobjective optimization algorithms, including, evolutionary algorithms.

Such multiobjective optimization algorithms provide a solution of the account's problem of the several objective functions (criteria, quality indicators) at the solution of various applied tasks.

Currently the genetic algorithms (GA) have the greatest application among the evolutionary multiobjective optimization algorithms. These algorithms have such advantages as:

- lack of the restrictions imposed on nature of the optimized objective functions;
- resistance to local optimum traps;
- high speed of convergence to the decision;
- solution's possibility of the tasks' wide class (including large-scale problems of optimization);
- simplicity of realization;
- use's possibility for the tasks with the changing environment.

Nowdays the following multiobjective optimization GA are famous and widely used:

- VEGA (Vector Evaluated Genetic Algorithm);
- FFGA (Fonseca and Fleming's Genetic Algorithm);
- NPGA (Niched Pareto Genetic Algorithm);
- NSGA (Non-dominated Sorting Genetic Algorithm);
- NSGA-II (Non-dominated Sorting Genetic Algorithm-II);
- SPEA (Strength Pareto Evolutionary Algorithm).

The VEGA [10] was proposed by D. Schaffer in 1984. It belongs to the group of the selection algorithms on the base of the switching objective functions. This algorithm is based on idea that use of the parents possessing the best variations of values of various objective functions (criteria, quality indicators), i.e. with their best sum (value of «supercriterion»), as a result will allow to receive the decision which will combine the best combination of values of various criterion functions (criteria, quality indicators) in total. This algorithm is the only algorithm among the presented algorithms which doesn't use the notion «Pareto set».

The FFGA [15] was proposed by Fonseca C.M. and Fleming P.J. in 1993. It applies the simplest variant of formation of Pareto set for selection of decisions.

Set of all decisions is ranged according to success of some decision with using of all objective functions (criteria, quality indicators) in relation to other decisions. The lower the rank of the solution, the greater probability of its choice for the next step of selection. The fundamental difference of the FFGA is that all objective functions (factors, quality indicators) are considered together (as a whole) in relation to the decision. Herewith not values of the objective functions, but ranks of all solutions in the population are analyzed.

The principal difference and advantage of the NPGA [12], which was proposed by Horn J., Nafpliotis N. and Goldberg D.E. in 1994, from the other genetic multiobjective optimization algorithms is that this algorithm has a mechanism to support the diversity of the solutions' population. This algorithm carries out a combination of the principles of tournament selection and the notation of Pareto dominance. The NPGA solves the problem of convergence to a local minimum, which is one of the main problems of the GA. There are various modifications of the NPGA. They differ in the approaches to the formation of the parent population.

The NSGA [13] was proposed by Srinivas N. and Deb K. in 1994. It applies some other approach to the problem's solution of the support of a variety in the parental population. For each generation not dominating sorting is also executed, but in addition, the so-called phenotype distance for the choice of decisions for selection is calculated. If this distance to any decision in the next generation is less, than is set by means of some threshold value, the considered decision isn't added to the next generation. However, this algorithm has some shortcomings. First, the ranging of decisions becomes superfluous if the phenotype distance is used. Secondly,

there is a need of determination of threshold value which sets admissible phenotype distance between the decisions.

The NSGA-II [14] was proposed by Deb K., Agrawal S., Pratap A. and Meyarivan T. in 2002. It provides correction of shortcomings of the NSGA. First, the improved sorting algorithm reduces computing complexity of calculations. Secondly, the decisions both from the modified set and from the initial set of decisions are used for formation of new population of decisions. So-called «crowding distance» is calculated for each decision. This distance allows to estimate, how some decision is close to the solutions-neighbors. Bigger mean value of «crowding distance» corresponds to the best variety of decisions in the population.

The SPEA [15] was proposed by Zitzler E. and Thiele L. in 1998. It uses the following approach. The decisions which aren't dominated by other decisions in the population are stored in the special external array. Thus the elitism mechanism which allows not to lose the good intermediate decision is realized.

The number of such selected decisions can be great. For reduction of the decisions' number which are stored in the external array, the clustering procedure is carried out. The SPEA effectively deals with the typical problem of the premature convergence which is often arising at realization of the principles of elitism. For this purpose the special mechanism of the niches' formation is used. Herewith the detection of the general suitability is carried out not on the base of distance between decisions, but on the base of the principles of Pareto dominance.

In the context of the problem's solution of development of the forecasting models on the base of the SBT and MCSA it is necessary to understand the forecasting model as the decision.

The analysis of literature devoted to the problems of multiobjective optimization with application of genetic algorithms demonstrates, that nowdays such algorithms as the SPEA, NSGA and NSGA-II find the greatest application at the solution of many applied tasks. So, for example, the NSGA and NSGA-II are successfully applied in the such problems as problem of scheduling, problem of drawing up of schedules, the travelling salesman problem [16]. Herewith the NSGA-II is significantly better than the NSGA because the NSGA-II minimizes the computing expenses.

In this regard the decision on expediency of adaptation of the ideas put in the NSGA-II at realization of the MCSA which is applied to selection of the forecasting models on the base of the SBT was made.

For confirmation of prospects of the offered transformation of the MCSA it is offered to realize the following algorithm of multiobjective optimization.

Step 1. To generate initial population of antibodies. Each antibody is coded on the base of the SBT and represents some forecasting model.

Step 2. To perform the nondominated sorting to population of antibodies on the base of two indicators of quality for the forecasting model (the affinity indicator (1) and the tendencies discrepancy indicator (2)).

Step 3. To choose the parents-antibodies for the next generation of the clones-antibodies based on the values of the rank and «crowding distance».

Step 4. To pass to step 5 if desirable values of the quality indicators are reached or the quantity of generations in the MCSA is settled. Otherwise to pass to step 2.

Step 5. To accept the antibody with the minimum value of the affinity indicator (1) in the last population as the optimum decision. To use the forecasting model corresponding to this antibody for forecasting.

As a result of application of the offered algorithm the Pareto set of the nondominated forecasting models will be received. These models provide the best combinations of values of the used quality indicators of the forecasting models for the analyzed TS.

The received forecasting models can be applied at the solution of a problem of medium-term forecasting. It will expand application scope of the forecasting models based on the SBT and MCSA.

3 Experimental studies

To confirm prospects of the offered algorithm, experimental studies on forecasting for 5 TS were conducted:

T1 (TS «The Brent crude oil price»; range of values: 15.12.2015 – 02.01.2016; unit of measure: ruble) = [46.04; 45.11; 44.46; 44.38; 44.30; 43.97; 44.70; 44.70; 44.08; 45.92; 47.08; 47.32; 47.46; 48.05; 48.84; 50.25; 48.90; 49.50];

T2 (TS «Total number of the made energy resources»; range of values: 1992 – 2012 years; unit of measure: quadrillion Btus) = [47.99; 44.69; 42.30; 41.42; 39.35; 38.74; 39.07; 40.81; 41.70; 42.63; 44.16; 47.16; 49.86; 51.05; 52.06; 52.52; 52.52; 50.01; 53.74; 54.63; 55.30];

T3 (TS «Number of Internet users per 100 people»; range of values: 1990 – 2014 years; unit of measure: percent) = [0.00000; 0.00067; 0.01345; 0.05382; 0.14815; 0.26975; 0.47296; 0.81274; 1.01899; 1.97723; 2.94437; 4.12827; 8.29886; 12.85939; 15.22667; 17.02328; 24.66000; 26.83000; 29.00000; 43.00000; 49.00000; 63.80000; 67.97000; 70.52000];

T4 (TS «Passenger traffic»; range of values: 1989 – 2012 years; unit of measure: millions passenger-kilometers) = [3669.202; 3635.441; 3305.227; 3145.413; 3087.016; 3251.615; 2714.314; 3268.434; 3743.247; 4249.406; 4468.123; 4489.569; 4651.419; 4726.832; 4986.468; 5054.200; 5482.314; 5655.296; 5129.103; 5987.268; 5933.041; 5973.308];

T5 (TS «Value added in agriculture»; range of values: 1992 – 2013 years; unit of measure: dollar) = [270112; 274000; 255000; 253200; 272475; 227102; 192117; 181200; 170300; 152900; 141042; 167100; 158000; 152900; 155573; 164272; 164262; 173699; 173411; 175800; 153500; 139028; 139842; 144612].

The chosen time series have different number of elements, the scale level of elements values, sampling points, amplitude and period. It allows to consider further conclusions universal.

Forecasting for each TS was executed with use of one (*Aff*) and two (*Aff* and *Tendency*) quality indicators of the forecasting model. The errors of forecasting are presented in Table 1.

Table 1. The erros of forecasting.

Error type	Forecasting error with use of *Aff* , %	Forecasting error with use of *Aff* and *Tendency* , %
TS #1 («The Brent crude oil price»)		
Training error (1)	3,12	1,10
Error for 1 step	1,43	0,22
Error for 2 step	0,85	0,01
Error for 3 step	0,67	0,38
Error for 4 step	1,06	0,22
Error for 5 step	0,45	0,57
Average error	0,89	0,28
TS #2 («Total number of the made energy resources»)		
Training error (1)	5,23	2,89
Error for 1 step	1,14	0,62
Error for 2 step	0,87	0,20
Error for 3 step	1,53	0,73
Error for 4 step	0,29	0,29
Error for 5 step	0,97	0,68
Average error	0,96	0,50
TS #3 («Number of Internet users per 100 people»)		
Training error (1)	5,49	3,29
Error for 1 step	1,48	0,57
Error for 2 step	3,75	0,13
Error for 3 step	7,41	0,65
Error for 4 step	5,86	0,12
Error for 5 step	9,14	0,42
Average error	5,53	0,38
TS #4 («Passenger traffic»)		
Training error (1)	5,47	2,54
Error for 1 step	1,58	1,75
Error for 2 step	2,84	1,42
Error for 3 step	3,54	1,98
Error for 4 step	3,17	1,10
Error for 5 step	3,69	1,38
Average error	2,96	1,52
TS #5 («Value added in agriculture»)		
Training error (1)	3,98	2,75
Error for 1 step	0,82	0,82
Error for 2 step	1,37	0,54
Error for 3 step	1,89	0,86
Error for 4 step	2,29	1,09
Error for 5 step	3,97	0,22
Average error	2,27	0,71

The errors of forecasting in the Table 1 are the average error values. These errors were calculated on the base of 1000 program runs for 100 generations of population. Size of population is set by 20 antibodies.

Values of the forward prediction errors (for 1 – 5 steps forward) specify that the offered approach to selection of the forecasting models is effective as for the solution of problems of short-term forecasting (for 1 – 3 step forward) as for the solution of problems of medium-term forecasting (for 4, 5 steps forward).

It should be noted that use of the additional quality indicator of the forecasting model (the indicator *Tendency*) allowed to carry out «search» of the forecasting model in the necessary (correct) direction. As a result, for all reviewed examples of time series for the small number of generations of the MCSA the smaller values of the affinity indicator *Aff* (training error (1)) and in most cases the smaller values of the forecasting errors for 1 – 5 steps forward were received (Table 1).

Results of forecasting for TS «The Brent crude oil price» have been presented in Figures 1 and 2. These results are received with use of the forecasting models on the base of one and two indicators of quality respectively.

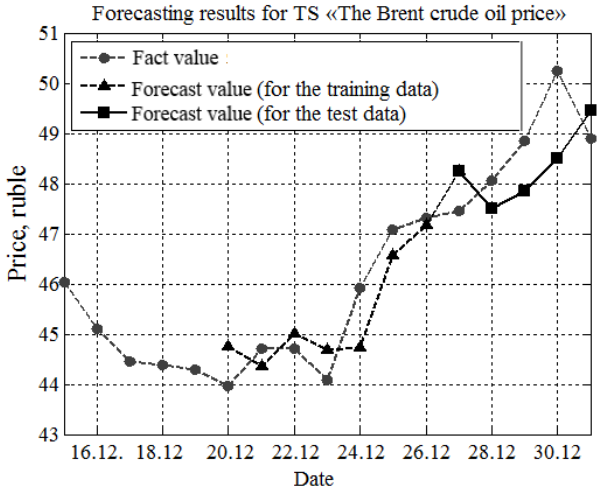

Figure 1. Forecasting of «The Brent crude oil price» (with one indicator of quality *Aff*).

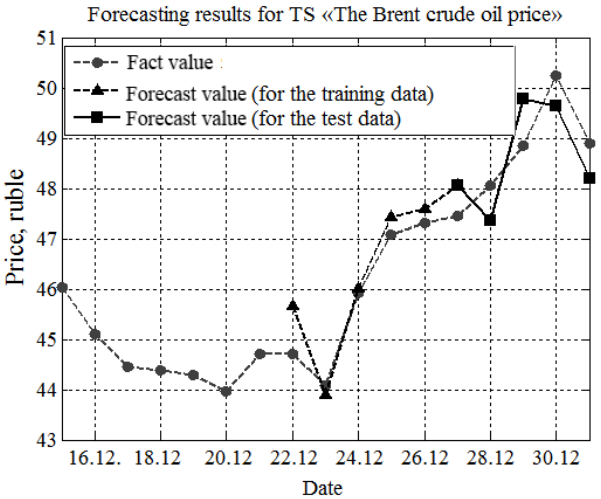

Figure 2. Forecasting of «The Brent crude oil price» (with two indicators of quality *Aff* and *Tendency*).

The presented graphic dependences show that the second model repeats the mathematical law of initial TS for the training set of data better than the first one. Moreover, this feature is kept also for the test set (when forecasting for 5 steps forward is carried out). Herewith the forecasting model on the base of two indicators of quality (*Aff* and

Tendency) demonstrates the best survival for the extended forecasting horizon than the forecasting model on the base of one indicator of quality (*Aff*). Besides this model is more effective for short-term forecasting.

These calculations allow to make the following conclusion: use of the second indicator of quality of the forecasting model provides a way to increase the life time of the forecasting model. Herewith the efficiency of the forecasting model for performing of the short-term forecasts saves that in general confirms the success of the offered approach.

4 Conclusions

Initially, the MCSA was developed for the solution of short-term forecasting problems. However, the fulfilled researches showed application's possibility of the MCSA for the solution of medium-term forecasting problems.

Apparently, application of the Pareto domination principles is the effective solution of the accounting problem of several quality indicators in the development problem of the forecasting models, which represent analytical dependences on the base of the SBT.

It should be noted that computing complexity of the MCSA, when the Pareto-optimal solutions are used, increases slightly. Herewith, it is possible to expand scope of application of the MCSA considerably.

Acknowledgment

This work is supported by Russian Federal Property Fund, 16-08-00771.

References

1. L.A. Demidova, Time series forecasting models on the base of modified clonal selection algorithm, *2014 International conference on computer technologies in physical and engineering applications (ICCTPEA),* pp. 33–34 (2014)
2. N.N. Astakhova, *Modern informatization problems in simulation and social technologies Proceedings of the XX-th International Open Science Conference,* pp. 146–150 (2015).
3. L.A. Demidova, Cloud Sci., **1**, 202 (2014)
4. L.A. Demidova, Automation and Remote Control, **74**, 313 (2013)
5. N.N. Astakhova, L.A. Demidova and E.V. Nikulchev, Contemporary Engineering Sciences, **8**, 1659 (2015)
6. N. Astakhova, L. Demidova, E. Nikulchev and E. Pluzhnik, *16th International Symposium on Advanced Intelligent Systems*, pp. 861–873 (2015).
7. N.N. Astakhova, L.A. Demidova and E.V. Nikulchev, Applied Mathematical Sciences, **9**, 4813 (2015)
8. N. Astakhova, L. Demidova and V. Konev, The Description Problem Of The Clusters' Centroids, *2015 International Conference "Stability and Control Processes" in Memory of V.I. Zubov (SCP)* , pp. 448–451 (2015).
9. Z. Michalewicz, *Proc. of the Sixth Int. Conf. on Genetic Algorithms and their Applications, Pittsburgh, PA,* pp. 239–247 (1995)
10. D.E. Goldberg, *Genetic Algorithms in Search, Optimization, and Machine Learning* (Addison-Wesley, 1989)
11. C.M. Fonseca and P.J. Fleming, *Multiobjective optimization and multiple constraint handling with evolutionary algorithms – Part I: A unified formulation, Technical report 564* (University of Sheffield, 1995)
12. J. Horn, N. Nafpliotis and D.E. Goldberg, *Proceedings of the First IEEE Conference on Evolutionary Computation,* **1**, 82 (1994)
13. J. Knowles and D. Corne, *Proceedings of the 1999 Congress on Evolutionary Computation,* pp. 98-105 (1999)
14. K. Deb, A. Pratap, S. Agarwal and T. Meyarivan, *A Fast and Elitist Multiobjective Genetic Algorithm: NSGA II.* (Indian Institute of Technology, 2000)
15. P.J. Bentley and J.P. Wakefield, *Proceedings of the 2nd On-Line World Conference on Soft Computing in Engineering Design and Manufacturing*, pp. 126–140 (1997)
16. K. Deb, *Multi-objective Optimization using Evolutionary Algorithms* pp. 221–232 (Wiley, 2001)