11. Лапко А. В., Лапко В. А. Синтез структуры семейства непараметрических решающих функций в задаче распознавания образов // Автометрия. 2011. Т. 47, № 4. С. 76–82.

12. Лапко А. В., Лапко В. А. Синтез структуры смеси непараметрических оценок плотности вероятности многомерной случайной величины // Системы управления и информационные технологии. 2011. № 1 (43). С. 12–15.

13. Лапко А. В., Лапко В. А. Анализ свойств непараметрических оценок смеси плотностей вероятности при различных условиях распределения статистических данных // Информатика и системы управления. 2013. № 1 (35). С. 119–126.

14. Лапко А. В., Лапко В. А. Непараметрические методики анализа множеств случайных величин // Автометрия. 2003. Т. 39, № 1. С. 54–61.

15. Лапко А. В., Лапко В. А. Регрессионная оценка плотности вероятности и ее свойства // Системы управления и информационные технологии. 2012. № 3 (49). С. 152–156.

16. Епанечников В. А. Непараметрическая оценка многомерной плотности вероятности // Теория вероятности и ее применения. 1969. Т. 14. Вып. 1. С. 156–161.

17. Лапко А. В., Лапко В. А. Оптимальный выбор количества интервалов дискретизации области изменения одномерной случайной величины при оценивании плотности вероятности // Измерительная техника. 2013. № 7. С. 24–27.

18. Lapko A. V., Lapko V. A. Optimal selection of the number of sampling intervals in domain of variation of a one-dimensional random variable in estimation of the probability density // Measurement Techniques. 2013. Vol. 56, no. 7. P. 24–27 (DOI: 10.1007/s11018-013-0279-x).

References

1. Lapko A. V., Lapko V. A. *Avtometriya*. 2002, no. 5, p. 38–48.

2. Lapko V. A., Kapustin A. N. *Avtometriya*. 2006, vol. 42, no. 6, p. 26–33.

3. Lapko A. V., Lapko V. A. *Avtometriya*. 2008, vol. 44, no. 3, p. 65–74.

4. Lapko A. V., Lapko V. A., Egorochkin I. A. *Sistemy upravlenija i informacionnye tehnologii*. 2009, no. 1 (35), p. 60–64.

5. Lapko A. V., Lapko V. A. *Sistemy upravlenija i informacionnye tehnologii*. 2009, no. 3.1 (37), p. 156–160.

6. Lapko A. V., Lapko V. A. *Avtometriya*. 2010, vol. 46, no. 1, p. 70–78.

7. Lapko A. V., Lapko V. A. *Sistemy upravlenija i informacionnye tehnologii*. 2010, no. 1 (39), p. 16–19.

8. Lapko A. V., Lapko V. A. *Sistemy upravlenija i informacionnye tehnologii*. 2010, no. 1.2 (39), p. 300–304.

9. Lapko A. V., Lapko V. A. *Sistemy upravlenija i informacionnye tehnologii*. 2010, no. 3 (41), p. 8–11.

10. Lapko A. V., Lapko V. A. *Avtometriya*. 2010, vol. 46, no. 6, p. 47–53.

11. Lapko A. V., Lapko V. A. *Avtometriya*. 2011, vol. 47, no. 4, p. 76–82.

12. Lapko A. V., Lapko V. A. *Sistemy upravlenija i informacionnye tehnologiju*. 2011, no. 1 (43), p. 12–15.

13. Lapko A. V., Lapko V. A. *Informatika i sistemy upravlenija*. 2013, no. 1 (35), p. 119–126.

14. Lapko A. V., Lapko V. A. *Avtometriya*. 2003, vol. 39, no. 1, p. 54–61.

15. Lapko A. V., Lapko V. A. *Sistemy upravlenija i informacionnye tehnologii*. 2012, no. 3 (49), p. 152–156.

16. Epanechnikov V. A. *Teorija verojatnosti i ee primenenija*. 1969, vol. 14, no. 1, p. 156–161.

17. Lapko A. V., Lapko V. A. *Izmeritel'naja tehnika*. 2013, no. 7, p. 24–27.

18. Lapko A. V., Lapko V. A. Optimal selection of the number of sampling intervals in domain of variation of a one-dimensional random variable in estimation of the probability density (2013) *Measurement Techniques*, 56 (7), p. 24–27. doi: 10.1007/s11018-013-0279-x.

УДК 004.93

INFORMATIVE ATTRIBUTE SELECTION WITH HYBRID SELF-ADJUSTED EVOLUTIONARY OPTIMIZATION ALGORITHM*

S. S. Volkova

Siberian State Aerospace University named after academician M. F. Reshetnev
31, Krasnoyarskiy Rabochiy Av., Krasnoyarsk, 660014, Russian Federation
E-mail: Svetlana.volkova.mail@yandex.ru

An informative attribute selection problem is considered. The problem is solved with the hybrid self-adjusted evolutionary algorithm. The algorithm is used as an optimization method of bandwidth parameters in kernel regression. The algorithm is experimented on the test function with various dimensions. Reliability depends on the dimension function which is also presented. The results of the hybrid self-adjusted algorithm are presented too.

Keywords: informative attribute selection, kernel regression, genetic algorithm, hybrid self-adjusted genetic algorithm.

* The study was supported by The Ministry of education and science of Russian Federation, project № 14.B37.21.1521.
The second International Workshop on Mathematical Models and its Applications (IWMMA 2013).

ОТБОР ИНФОРМАТИВНЫХ ПРИЗНАКОВ С ПОМОЩЬЮ ГИБРИДНОГО САМОНАСТРАИВАЮЩЕГОСЯ ЭВОЛЮЦИОННОГО АЛГОРИТМА ОПТИМИЗАЦИИ

С. С. Волкова

Сибирский государственный аэрокосмический университет имени академика М. Ф. Решетнева
Российская Федерация, 660014, г. Красноярск, просп. им. газ. «Красноярский рабочий», 31
E-mail: Svetlana.volkova.mail@yandex.ru

Рассматривается задача отбора информативных признаков. Задача решается с помощью гибридного самонастраивающегося эволюционного алгоритма, используемого в качестве метода оптимизации для определения оптимальных значений параметров размытия непараметрической оценки регрессии. Представлены исследования эффективности применения рассматриваемого алгоритма на тестовых задачах различной размерности. Показаны полученные зависимости эффективности применения рассматриваемого метода от размерности задач. Продемонстрирована эффективность предлагаемого подхода к самонастройке эволюционных алгоритмов.

Ключевые слова: отбор информативных признаков, непараметрическая оценка регрессии, генетический алгоритм, гибридный самонастраивающийся генетический алгоритм.

1. Introduction

Processes in technical and organization systems are described with many attributes. But using of non-informative attributes in model is negative for research. And so processing with real data should start with selection of informative attributes.

The informative attribute selection problem is well-known. There are many classical methods for solving this problem (Factor analysis methods, Principal component analysis) [1]. But these methods usually have limitations in practice, i. e. principal component analysis can be not appropriate for processes with strong nonlinear relations among variables. And now new methods are investigated.

Informative attribute selection problem is solved by regression models in this paper. They base on Nadaraya-Watson nonparametric estimation (kernel regression) [2]. This estimation has advantages: the method does not need to search a structure of a mathematical model of the process and it can be applied for processes with nonlinear relations among variables.

Nonparametric regression estimation (Nadaraya-Watson kernel regression) depends on parameter "bandwidth" (see section 2). And so successful at using of kernel regression comes to optimization problem. If kernel regression is used for multivariable function, it is necessary to select bandwidth for each variable. Using of classical optimization methods is difficult. This problem has non-analytical form and high dimension. One of the possible methods for this optimization problem may be evolutionary algorithm [3]. In paper it is proposed to use genetic algorithm. In this case individuals are binary code representation of the vector of bandwidth parameters at nonparametric regression estimation. Fitness function is average error of the nonparametric regression estimation for test sample.

Idea of informative attribute selection bases on one of the properties of non-parametric evaluation. Bandwidths of non-informative attributes trend to high values (see section 2). And so high optimal (suboptimal) value of bandwidth means low information content of the relevant attribute.

It is necessary to notice that genetic algorithm has disadvantage because reliability of genetic algorithm greatly depends on algorithm settings. And experienced researcher need to spend a lot of time to select effective settings, so it is necessary to use self-adjusted algorithm. In paper it is proposed to use hybrid self-adjusted genetic algorithm.

This work is outlined as follows: Information about kernel regression is presented in section 2. Information about genetic algorithm and algorithm setting is introduced in section 3. Experiments are described in section 4; results of algorithm running are discussed in section 4 too.

2. Kernel regression

Let's us consider Nadaraya-Watson kernel regression. The kernel regression is a non-parametric technique in statistics to estimate the conditional expectation of a random variable.

Let $(x_1, x_2, \ldots, x_n)$ be a vector of variable values, y be a regression value, N be a training sample size. Non-parametric regression estimation for variable vector $(x_1^*, x_2^*, \ldots, x_n^*)$ looks like (1):

$$\hat{y}(\overline{x}^*) = \frac{\sum_{i=1}^{N} y_i \cdot \prod_{j=1}^{n} \Phi\left(\frac{x_j^i - x_j^*}{c_j}\right)}{\sum_{i=1}^{N} \prod_{j=1}^{n} \Phi\left(\frac{x_j^i - x_j^*}{c_j}\right)}, \qquad (1)$$

here c_j is bandwidth parameter, $\Phi(\ldots)$ is kernel function (weighting function). Quality of regression estimation does not depend on selection of kernel function significantly. This choice is determined by requirements of estimation differentiability. Triangular is one of the common types of kernel functions:

$$\Phi(t) = \begin{cases} 1 - |t|, & \text{if } |t| < 1, \\ 0, & \text{else.} \end{cases}$$

Quality criterion is used for assessment regression estimation. In paper quality criterion is mean-square error (2):

$$W = \frac{1}{N_t}\sum_{i=1}^{N_t}\left(\hat{y}_i - y_i\right)^2, \quad (2)$$

here N_t is an exam sample, y_i is regression value (from exam sample), $\hat{y}_i$ is estimation of regression value (calculated by kernel regression).

Problem of non-parametric regression estimation construction comes to selection of the best values of bandwidth parameters c_j, i. e. to minimize quality criterion W with bandwidth parameters c_j.

It is necessary to notice optimal bandwidth parameter c_j is increasing for non-informative attributes [4]. If c_j trends to infinity ($c_j \to \infty$), argument of kernel function $\Phi(x_j)$ trends to zero, and $\Phi(0) = 1$. In this situation kernel regression (1) does not depend on variable value x_j.

3. Hybrid self-adjusted genetic algorithm for optimization of bandwidth parameter

Optimization of the bandwidth parameters is complicated problem because the problem has non-analytical form and dimension can be high. So an appropriate tool for such problem solving can be genetic algorithm.

Genetic algorithm is a stochastic optimization method of direct search. And so this algorithm is wide applicability. And genetic algorithm can process with complex discontinuous implicitly defined function. This optimization method copies evolution processes and uses idea of collective learning. Population is a set of individuals. Each individual is point of search space and takes part in collective learning. Individual has fitness calculated with value of criterion function (quality criterion). Genetic algorithm has stages such as generation formation, selection, crossover, and mutation. Individuals compete with other individuals in the population for transmission of its genetic information to the next population (selection stage). Selected individuals are parents for creating new offspring-individuals (crossover stage). In our case individuals are binary code representation of the vector of bandwidth parameters at nonparametric regression estimation. Fitness function is average error of the nonparametric regression estimation for train sample.

There are various selection types (proportional, rank, tournament), various crossover types (two-point, one-point, uniform), and various mutation types (week, average, strong). It means there are many combinations of algorithm settings.

Genetic algorithm greatly depends on algorithm settings. There are no universal settings because genetic algorithm realizes two strategies. First strategy is exploration. Aim is search of new solution areas. It is the most valid on initial stages of search. Second strategy is execution. It needed for improving of existing decision. It is the most valid on final stages of optimization algorithm. In genetic algorithm mutation realizes exploration strategy, crossover realizes execution strategy. Also the most effective algorithm settings depend on function topology. And so it is necessary to implement algorithm with self-adjusted settings during optimization process.

Self-adjusted method is based on hybrid self-adjusted evolutionary Gomez algorithm. This algorithm combines genetic algorithm and evolutionary strategies [5]. So it names hybrid method. The idea of the algorithm is as follows: every individual has personal probability of using each type of genetic operators. The algorithm randomly selects combination of settings (proportional selection) and runs with each individual in population with personal settings. In the end of generation the offspring is compared with its parent. If fitness function value of the offspring is better than fitness function value of the parent, then probability of selected operator types are increased. Otherwise probabilities of selected operator types are decreased. Probability of using selected operator is evaluated with (3):

$$\begin{array}{l} p_k = (1+\delta)\cdot p_k, k = 1,\ 2,\ 3, \\ if\ (fitness(offspring) \geq fitness(ind_i)); \\ p_k = (1-\delta)\cdot p_k, k = 1,\ 2,\ 3, \\ if\ (fitness(offspring) < fitness(ind_i)), \end{array} \quad (3)$$

here k is sequence number of operator (selection, crossover, mutation); p_k is probability of using selected type operator; δ is a training parameter that is generated with equally probable distribution distribution on interval [0, 1]; *offspring* is new individual ind_i is parent-individual;

fitness(ind) is fitness function value of individual *ind*.

After that probabilities of all types of all genetic operators are normalized. It is necessary to notice in first population the probability of using various genetic operators is the same.

So we use hybrid self-adjusted evolutionary optimization algorithm for bandwidth parameters optimization at nonparametric estimation regression.

4. Experiments and Result

Linear combinations of input attributes are taken as test functions for our method:

1) $y(\bar{x}) = 0,01\cdot x_1 + 7\cdot x_2 + 5\cdot x_3$;
2) $y(\bar{x}) = 0,01\cdot x_1 + 7\cdot x_2 + 5\cdot x_3 + 12\cdot x_4 + 8\cdot x_5$;
3) $y(\bar{x}) = 0,01\cdot x_1 + 7\cdot x_2 + 5\cdot x_3 + 12\cdot x_4 + 8\cdot x_5 + 15\cdot x_6 + 3\cdot x_7$;
4) $y(\bar{x}) = 0,01\cdot x_1 + 7\cdot x_2 + 5\cdot x_3 + 12\cdot x_4 + 8\cdot x_5 + 15\cdot x_6 + 3\cdot x_7 + 9\cdot x_8 + 13,5\cdot x_9$;

Functions with various dimensions are examined. And all functions have non-informative attribute (with low weight coefficient). The aim of research is to identify non-informative attributes.

Training sample (100 points) is generated randomly from the interval [0, 3] with equally probable distribution for each attributes. Test sample (100 points) is generated randomly too for the same interval. Algorithm is tested without noise and with noise (10 %). We use centered noise with mean equals zero and we add it randomly with

with equally probable distribution. Algorithm has 50 individuals for 50 generations. We tested each setting of the genetic algorithm for each problem for 100 times. After that we calculated reliability of the algorithm. It means percentage of the successful runs of the algorithms. Successful run means that algorithm found the least important variable with the highest value of the bandwidth parameter. We have got values of reliability for 10 times for statistical significance of our numerical experiments and corresponding conclusions.

Experiments were conducted at the same algorithm resource on function with various dimensions because it is necessary to estimate effectiveness of self-adjusted genetic algorithm in the same conditions and fall of algorithm reliability with increasing dimensions.

Implemented genetic algorithm runs 20 times for each combination of settings. After that bandwidth parameters are averaged for each attributes. In each runs the algorithm finds non-informative attributes, mean-square error of kernel regression with all attributes and without each attribute is calculated. Obtained statistical data is processed; results are presented in tables and graphs.

Previously numerical research was investigated with all combinations of algorithm settings. Genetic algorithms with different settings have various reliabilities on test functions [6]. Results were averaged like expectation value, if researcher does not know the best algorithm settings. The most effective algorithm settings were found with exhaustive search, and result on these settings was saved. Fig. 1 and 2 display averaged results and result on the best settings and with self-adjusted result.

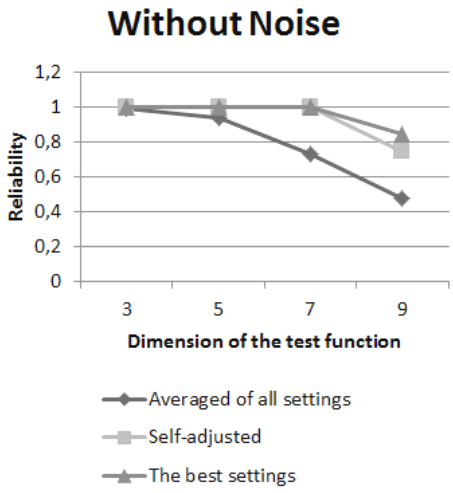

Fig. 1. Algorithm reliability dependence of test function dimension without noise

Obviously reliability of hybrid self-adjusted genetic algorithm above averaged results and little below result on the best settings. And so if researcher does not know the most effective algorithm settings, he should use self-adjusted algorithm. Hybrid self-adjusted genetic algorithm does not increase running time compared with the standard algorithm because much time is required to calculate the fitness function and these algorithms have the same number of fitness function calculations on the generation.

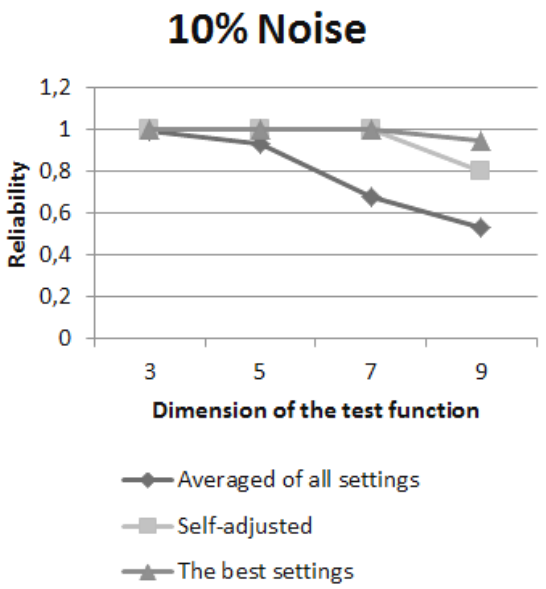

Fig. 2. Algorithm reliability dependence of test function dimension with 10 % noise on training sample

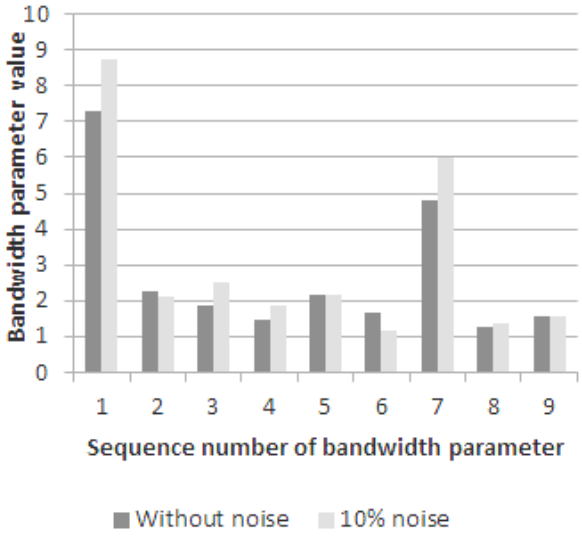

Fig. 3. Bandwidth parameter values for test function 4

You should pay attention to research result like result of informative attribute selection problem with kernel regression [6]. Fig. 3 and table present results of genetic algorithm on test function 4. Fig. 3 shows the values of bandwidth parameters. It is necessary to notice bandwidth parameter values are increased with reducing of weight of corresponding attribute in criterion function value. That is to say if you make decreasing sequence of bandwidth value, attributes is located mainly in order of increasing attribute weight. Only mean-square error of kernel regression without non-informative attribute may compare with mean-square error with all attributes (see table).

Fig. 3 and table show the final results for the different features with the conclusion that features 1 and 7 are the least important ones.

Mean-square error of kernel regression on test function 4

	Without noise	10 % noise
Mean-square error without attribute 1	96,7	128,38
Mean-square error without attribute 2	136,43	148,63
Mean-square error without attribute 3	107,71	144,03
Mean-square error without attribute 4	185,31	221,83
Mean-square error without attribute 5	128,82	152,7
Mean-square error without attribute 6	261,45	261,23
Mean-square error without attribute 7	98,91	129,77
Mean-square error without attribute 8	144,75	179,09
Mean-square error without attribute 9	190,72	228,52
Mean-square error with all attributes	98,07	127,98

In addition we tested parallel version of our method for multiprocessor based on the parallelization of the genetic algorithm (fitness function calculation). We used multiprocessor with 2, 4, and 6 cores. And we got following values of the computing speed-up coefficients:

1) Configuration with 2 cores: 1,81–1,85;
2) Configuration with 4 cores: 3,20–3,76;
3) Configuration with 6 cores: 3,63–4,14.

So we can conclude that our method is appropriate for parallelization and using for multiprocessors.

5. Conclusions

So hybrid self-adjusted evolutionary algorithm is implemented for informative attribute selection. Reliability of its algorithm was experimented.

Genetic algorithm effectively solves optimization problem with bandwidth parameters in kernel regression. Hybrid self-adjusted algorithm solves problem of algorithm setting.

So hybrid self-adjusted genetic algorithm solves informative attribute selection problem on test functions effectively. Also this algorithm gives some data for analysis of information content of the attributes. The method is appropriate for parallelization.

References

1. Aivazyan, S. A. [et al.] Applied statistics. Classification and reduction of dimensionality. Moscow, Finansy i statistika, 1989, 607 p.

2. Medvedev A. V. Non-parametrical systems of adaptation. Novosibirsk, Nauka, 1983. 174 p.

3. Goldberg D. E. Genetic algorithms in search, optimization and machine learning. Reading, MA: Addison-Wesley, 1989.

4. Hall P., Li Q. and J. S. Racine, «Nonparametric Estimation of Regression Functions in the Presence of Irrelevant Regressors. *Review of Economics and Statistics*. 2007. 89. P. 784–789.

5. Gomez J. Self-adaptation of operator rates in evolutionary algorithms. *Proc. of Genetic and Evolutionary Computation Conference.* 2004. P. 1162–1173.

6. Volkova S., Sergienko R. B. Informative attributes selection in non-parametric regression estimation by making use of genetic algorithms. *Vestnik Tomskogo gosudarstvennogo universiteta. Ypravlenie, vichislitelnaya technika i informatica*. 2013. № 1 (22). P. 40–48.

УДК 004.6

РАЗВИТИЕ МЕТОДОВ ЭКВИВАЛЕНТНОГО ПРЕОБРАЗОВАНИЯ ГЕРТ-СЕТЕЙ ДЛЯ АНАЛИЗА МУЛЬТИВЕРСИОННОГО ПРОГРАММНОГО ОБЕСПЕЧЕНИЯ

Д. И. Ковалев, М. В. Сарамуд, М. В. Карасева, Ю. А. Нургалеева

Сибирский государственный аэрокосмический университет имени академика М. Ф. Решетнева
Российская Федерация, 660014, г. Красноярск, просп. им. газ. «Красноярский рабочий», 31
E-mail: saramud@bk.ru

В настоящее время практически все большие программные системы являются распределенными. Анализ мультиверсионного программного обеспечения (ПО) можно проводить, основываясь на распределенных системах обработки информации. На примере ГЕРТ-сети, моделирующей поведение системы Condor при расчете задачи с фиксированной продолжительностью в режимах без резервного копирования и миграции (режим Vanilla), показаны эквивалентные преобразования, позволяющие существенно упростить сеть и облегчить поиск петель. Приведены расчеты, позволяющие получить вероятностные характеристики приведенной сети. Описываются различные режимы работы системы Condor, преимущества распределенных гетерогенных систем обработки информации.

Ключевые слова: мультиверсионное программное обеспечение, ГЕРТ-сети, вероятностные характеристики.