VARIANTS OF ENCODING FOR SELECTION OF OPTIMAL SUBSET OF DIAGNOSTIC TESTS

Anna Yankovskaya, Yury Tsoy

Abstract: The paper concerns problem of selection of optimal subset of irredundant unconditional diagnostic tests by means of evolutionary approach. Three different variants of genetic encoding to solve this problem are described. Also new view on the optimal tests subset selection problem considering multi-objective variant of the well-known traveling salesman problem is introduced. The suggestion is made that evolutionary programming approach would be more appropriate then genetic algorithm because of disadvantage of crossover use for multi-objective problems solution.

Keywords: optimal tests subset selection, evolutionary multi-objective optimization, diagnostic test, intelligent systems, coevolution, genetic encoding.

ACM Classification Keywords: G.1.6 [Mathematics of Computing]: Optimization – Constraint optimization

Introduction

Selection of "good" irredundant unconditional diagnostic tests (IUDT) is of great importance for decision making in intelligent systems, since quality of obtained solutions depends significantly on properties of the used tests. However such a selection doesn't necessarily lead to an optimal solution because total number of features in selected tests set can be too large as well as time consumption and cost. Also one should take into consideration damage (risk), caused in result of features measuring for the object under investigation, for example, in geoecological or biomedicine problems.

This research continues our previous work on optimal subset of IUDTs selection [Yankovskaya, 2002, Yankovskaya, Mozheiko, 2004, Kolesnikova et al., 2005, Yankovskaya, Tsoy, 2005]. For the first time the optimization criteria and the problem of optimal tests subset selection has been formulated in the paper [Yankovskaya, 2002]. In the paper [Yankovskaya, Mozheiko, 2004] logical-combinatorial algorithm for optimal IUDTs subset selection was presented. In the paper [Kolesnikova et al., 2005] optimization criteria were further elaborated and three algorithms providing satisfaction of those criteria were proposed: logical-combinatorial with sequent satisfaction of the prescribed criteria, algorithm of optimal tests set selection on the base of hierarchies analysis method, and genetic algorithm (GA).

For solution of the optimal IUDT subset selection problem we will use evolutionary algorithm (EA) which presents heuristic search concept similar to "trials-and-errors" method. In this paper we propose two new variants of genetic encodings for candidate-solutions and also present another view on the formulated problem introducing multi-objective free traveling salesman problem – MOFTSP.

During last decade a number of models of GAs were developed, such as NSGA-II [Deb et al., 2002], PAES [Knowles, Corne, 2000], SPEA2 [Zitzler, Laumanns, Thiele, 2001], PPREA [Hallam, Graham, Blanchfield, 2006] to solve multi-objective optimization (MOO) problems. Also alternative approaches on a basis of particles swarm optimization [Alvarez-Benitez, Everson, Fieldsen, 2005] and differential evolution [Becerra, Coello Coello, 2006] were proposed. Some researches are aimed at reduction of the optimization criteria number (see for example [Brockhoff, Zitzler, 2006]) and this certainly appears to be promising for the optimization results, though search of competent universal method of reduction of criteria number is rather challenging (if ever possible) due to great variety of existing MOO problems.

One of the critical conditions for the success of EA in MOO problem is preserving as many undominated (incomparable) solutions within one population as possible. Such solutions correspond to different points on the Pareto front. To preserve population of undominated solutions an idea of grouping of individuals according to some similarity/difference measure emerges in various forms, for example, as niching, or as specific non-dominated selection [Deb et al., 2002]. Considering this condition the idea of GAs use to solve MOO problems looks rather contradictory, from the authors point of view, since the main searching operator in GA is crossover and use of this operator traditionally involves risk of recombination of incompliant values of the optimization parameters due to crossing of different parent individuals, though the last can be situated rather close to each other in parameters space.

We are planning to examine this by investigation of MOO problem solution using evolutionary programming (EP) algorithm, which doesn't adopt crossing of individuals. The results of EP optimization will be compared with those of GA.

Basic Notions and Definitions

Let's introduce a number of definitions [Yankovskaya, 2002, Yankovskaya, Mozheiko, 2004, Yankovskaya, 1996] and notations used in this paper.

Test is a set of features distinguishing any pair of objects belonging to different patterns.

The test is called *irredundant* if after the removal of any feature the test is not a test.

The feature is called *obligatory* if it is contained in all irredundant tests [Yankovskaya, 2000].

The feature is called *pseudoobligatory* if it is not obligatory and enters the set of irredundant tests used in decision making.

Let $\mathbf{T} = \{t_{ij} \mid i = 1,...,n,\ j = 1,...,m\}$ be the matrix of IUDTs and $\mathbf{T}_i$ corresponds to the i^{th} IUDT (the i^{th} row of matrix $\mathbf{T}$). We denote set of characteristic features as $\mathbf{z} = \{z_j \mid j = 1,...,m\}$ and for each feature z_j we define its weight w_j [Yankovskaya, 1996], cost w'_j [Yankovskaya, Mozheiko, 2004] and damage w''_j [Yankovskaya, Tsoy, 2005].

The case of binary matrix $\mathbf{T}$ is considered therefore the weight W_i of the i^{th} IUDT is

$$W_i = \sum_j w_j t_{ij} \ .$$

Then average test weight along all tests inside the IUDT matrix equals to:

$$\overline{W} = \frac{\sum_i W_i}{n} \ .$$

Number η_i of features in each test is given by $\eta_i = \sum_j t_{ij}$ and average number of features along all tests in $\mathbf{T}$ is

$$\overline{\eta} = \frac{\sum_i \eta_i}{n} \ .$$

Setting of a Problem

For the given tests matrix $\mathbf{T}$ with defined values of features weight, cost and damage it is necessary to find such submatrix $\mathbf{T}_0$ with n_0 rows, which corresponds to the set $\mathbf{N}^0$ of tests that would provide satisfaction of the following criteria (in order of significance descend):

1. $\mathbf{N}^0$ should contain as many pseudo-obligatory features as possible.
2. $\mathbf{N}^0$ should contain in total as small number features as possible.
3. $\mathbf{N}^0$ should have maximum possible total weight.
4. $\mathbf{N}^0$ should have minimum possible total cost.
5. $\mathbf{N}^0$ should have minimum possible total damage.

Statement of this problem accounting 5 optimization criteria was firstly introduced in the paper [Kolesnikova et al., 2005]. Since solving of the problem at hand is considered with use of evolutionary algorithm, which is known to be a heuristic search method, then as a consequence there is no guarantee that the optimal submatrix $\mathbf{T}_0$ (subset of IUDTs) will be found. In other words obtained solution is most likely to be suboptimal.

The problem formulated in this Section can also be considered as a modification of the well-known traveling salesman problem but here salesman is traveling for free and can visit only n_0 cities (not all the n ones) and in each city he has definite income (from sales) and expenses (cost of staying in the city). The task is to find such a path which provides the largest total income and the least total expenses. We will refer to this problem formulation as a multi-objective free traveling salesman problem – MOFTSP.

Genetic Encodings

We are going to use for comparison the following encoding schemes (for example shown on fig. 1a):

1. Candidate-solutions are encoded in binary chromosomes (strings) of length n, where each i^{th} symbol denotes inclusion ("1") (exclusion ("0")) of the i^{th} IUDT in (from) the resulting set of tests (fig. 1b). Note that number of units in the chromosome (number of IUDTs included in the resulting subset) can be unequal to n_0 therefore an additional constraint should be added for control of the number of units in chromosomes.

$$\mathbf{T} = \begin{array}{c|cccccccccccc|} & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 \\ 1 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 1 \\ 2 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 3 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 1 & 0 & 0 & 1 \\ 4 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 1 & 0 & 0 \\ 5 & 1 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 1 \\ 6 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 0 \\ 7 & 0 & 1 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 1 \end{array}$$

$$\mathbf{T}_0 = \begin{array}{c|cccccccccccc|} & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 \\ 2 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 4 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 1 & 0 & 0 \\ 6 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 0 \end{array}$$

a) Initial matrix $\mathbf{T}$ and solution submatrix $\mathbf{T}_0$

b) Solution representation for the 1[st] encoding scheme:	{0,1,0,1,0,1,0}
c) Solution representation for the 2[nd] encoding scheme:	{2,4,6}
d) Solution representation for the 3[rd] encoding scheme:	{0,1,0,1}U{0,1,0}

Fig 1. Example of solution representation for different encoding schemes.

2. In case of the 2nd encoding scheme each chromosome includes n_0 integer-coded parameters where each parameter corresponds to the ordinal number of the IUDT in the initial matrix $\mathbf{T}$ (fig. 1c). In the case of this encoding scheme each chromosome should contain only distinct (mutually unequal) values of parameters. From the MOFTSP viewpoint the salesman should not come twice (or more) to the one and the same city.

3. The 3rd encoding scheme uses cooperative coevolution idea [Potter, De Jong, 2000]. There are several subpopulations. Each one deals with its range of rows (submatrix of $\mathbf{T}$) such that submatrices for different subpopulations do not overlap. Chromosomes for each subpopulation are considered as binary strings analogous to the 1st encoding scheme. The candidate-solution is constructed by concatenation of the representative chromosomes from different subpopulations resulting in the binary chromosome similar to the chromosome for the 1st encoding scheme (example for the case of 2 subpopulations where the 1st one deals with rows 1-4 and the 2nd – with rows 5-7, is shown in fig 1d).

Let's make some comments on encodings under use.

First of all note that in case of use of the 1st and the 3rd encoding schemes there is additional optimization constraint with the greatest weight. Therefore we can expect that certain number of generations in the beginning of the evolutionary search will be spend to find the candidate-solutions that correspond to the IUDTs subset of power n_0. The search of the solution satisfying to the prescribed optimization criteria can be performed only when this stage is over. In this connection search time for the case of the 1st and the 3rd encoding schemes is expected to be larger than that of for the 2nd encoding case. To overcome this deficiency of the 1st and the 3rd encodings an initialization of the binary chromosomes including exactly n_0 units can be proposed.

Use of the 2nd encoding scheme is connected with the problem mentioned above in this section. Since no IUDT can be included twice or more in the resulting subset, there should be a mechanism that eliminates incorrect candidate-solutions. Next, note that enumeration order of the numbers of tests included at the resulting subset doesn't matter. In other words, permutations of parameters inside the chromosome doesn't change the result (since the salesman is traveling for free). For example, solution shown in fig. 1c can also be presented as {2,6,4} or {6,2,4} etc. Such an uncertainity involves the probability of presenting inside the population different permutations of the one and the same candidate-solution and thus slows the evolutionary search. To avoid this we will sort parameters inside the chromosome in the increasing order.

Objective Function

We will calculate fitness of the individual with chromosome h by evaluation of quality of corresponding submatrix $\mathbf{T}(h)$ as follows [Yankovskaya, Tsoy, 2005]:

$$f_h = \sum_{k=1}^{5} v_k e_h^{(k)} + 100\big(U(h) - n_0\big)^2,$$

where v_k is a weight coefficient for the kth optimization criterion corresponding to its significance; $U(\boldsymbol{\psi})$ gives number of units in binary string $\boldsymbol{\psi}$; $e_h^{(k)}$ is a penalty function for violation of the kth criterion:

$$e_h^{(1)} = \frac{m - U_c\big(\mathbf{T}_0(h)\big)}{m}, \quad e_h^{(2)} = \frac{U_d\big(\mathbf{T}_0(h)\big)}{m},$$

$$e_h^{(3)} = \frac{S_W(\mathbf{T}) - S_W(\mathbf{T}_0(h))}{S_W(\mathbf{T})}, \quad e_h^{(4)} = \frac{S_{W'}\big(\mathbf{T}_0(h)\big)}{S_{W'}(\mathbf{T})},$$

$$e_h^{(5)} = \frac{S_{W''}\left(\mathbf{T}_0(h)\right)}{S_{W''}(\mathbf{T})},$$

where $S_W(\mathbf{\Psi}), S_{W'}(\mathbf{\Psi})$ and $S_{W''}(\mathbf{\Psi})$ – total weight, cost and damage correspondingly along all tests of the set of IUDTs corresponding to matrix $\mathbf{\Psi}$; $U_c(\mathbf{\Psi}) = U\left(\bigwedge_i \mathbf{\psi}_i\right)$ and $U_d(\mathbf{\Psi}) = U\left(\bigvee_i \mathbf{\psi}_i\right)$ – correspondingly number of units in conjunction and disjunction along all rows of binary matrix $\mathbf{\Psi}$. Evolutionary search is aimed at minimization of f.

In order to respect priorities of criteria mentioned above we will reduce weights of penalties with growth of penalty number *k*. Then the following penalties weights will be used: $v_1 = 40$, $v_2 = 30$, $v_3 = 15$, $v_4 = 10$, $v_5 = 5$. Note that penalties weights depend on the specific application.

Conclusion

Three variants of genetic encoding schemes to solve problem of optimal tests subset selection had been introduced in this paper. Also a new variant of the problem under consideration: the multi-objective free traveling salesman problem – MOFTSP had been introduced. It's worth noting that the optimal tests subset selection problem can also be reduced to a problem of search of optimal row coverings for Boolean matrix [Yankovskaya, Gedike, 1999].

In result of critical analysis of application of GA for solution of MOO problems and suggested deficiencies involved by crossover operator, use of EP algorithm instead of GA is proposed.

Future work is connected with experimental comparison of use of GA and EP with different encodings for the solution of the formulated problem of optimal IUDTs subset selection.

Implemented algorithms will be used in instrumental intelligent tool IMSLOG [Yankovskaya et al., 2003] for regularities revealing and decision making on the basis of test pattern recognition.

Bibliography

[Yankovskaya, 2002] A.E. Yankovskaya. Construction of logical tests of prescribed properties and logic-combinatorial pattern recognition on them // Abstracts of the Intellectualization of Information Processing (IIP2002). Simferopl, 2002. P. 100-102. (in Russian).

[Yankovskaya, Mozheiko, 2004] A.E. Yankovskaya, V.I. Mozheiko. Optimization of a set of tests selection satisfying the criteria prescribed // 7th International Conference on Pattern Recognition and Image Analysis: New Information Technologies (PRIA-7-2004). Conference Proceedings. Vol. I. – St. Petersburg: SPbETU 2004. – Pp.145-148.

[Kolesnikova et al., 2005] S.I. Kolesnikova, V.I. Mozheiko, Y.R. Tsoy, A.E. Yankovskaya. Algorithms of selection of optimal set of irredundant diagnostic tests in intelligent decision making systems // Proceedings of the First International conference "System analysis and information technologies" (SAIT-2005). Vol. 1. Moscow: KomKniga, 2005. P. 256-262.

[Yankovskaya, Tsoy, 2005] A.E. Yankovskaya, Y.R. Tsoy. Optimization of a set of tests selection satisfying the criteria prescribed using compensatory genetic algorithm // Proceedings of IEEE East-West Design & Test Workshop (EWDTW'05) Odessa, Ukraine, September 15-19, 2005. Kharkov: SPD FL Stepanov V.V. P. 123-126.

[Deb et al., 2002] K. Deb, S. Agrawal, A. Pratap, T. Meyarivan. A fast and elitist multi-objective genetic algorithm: NSGA-II. IEEE Transactions on Evolutionary Computation, 2002, vol. 6, no. 2, pp. 182–197.

[Yankovskaya, 2000] Yankovskaya A.E. Logical Tests and Cognitive Graphic Means in intelligent system // New Information Technologies in Investigations of Discrete Structures: Proceedings of the 3-d All-Russian Conf. with Foreign Participants. – Tomsk: SO RAS, 2000. – P. 163-168. (in Russian).

[Knowles, Corne, 2000] J.D. Knowles, D.W. Corne. Approximating the non-dominated front using the Pareto archived evolution strategy. Evolutionary Computation Journal, 2000, vol. 8, no. 2, pp. 149–172.

[Zitzler, Laumanns, Thiele, 2001] E. Zitzler, M. Laumanns, L. Thiele. SPEA2: Improving the strength pareto evolutionary algorithm for multi-objective optimization. InK. C. Giannakoglou, D. T. Tsahalis, J. P'eriaux, K. D. Papailiou, and T. Fogarty, editors, Evolutionary Methods for Design Optimization and Control with Applications to Industrial Problems, pages 95–100, Athens, Greece, 2001.

[Hallam, Graham, Blanchfield, 2006] N. Hallam, K. Graham, P. Blanchfield. Solving multi-objective optimization problems using the potential pareto regions evolutionary algorithm // T.P. Runarsson et al. (eds.): PPSN IX, LNCS 4193, pp. 503-512. Springer-Verlag, 2006.

[Alvarez-Benitez, Everson, Fieldsen, 2005] J.E. Alvarez-Benitez, R.M. Everson, J.E. Fieldsen. A MOPSO algorithm based exclusively on pareto dominance concepts // In C. Coello-Coello et al. (eds.): Evolutionary Multi-Criterion Optimization, vol. 3410, pp. 459-473. Springer, 2005.

[Becerra, Coello Coello, 2006] R.L. Becerra, C.A. Coello Coello. Solving hard multi-objective optimization problems using -constraint with cultured differential evolution // T.P. Runarsson et al. (eds.): PPSN IX, LNCS 4193, pp. 543-552. Springer-Verlag, 2006.

[Brockhoff, Zitzler, 2006] D. Brockhoff, E. Zitzler. Are all objectives necessary? On dimensionality reduction in evolutionary multi-objective optimization // T.P. Runarsson et al. (eds.): PPSN IX, LNCS 4193, pp. 533-542. Springer-Verlag, 2006.

[Yankovskaya, 1996] A.E. Yankovskaya. Design of Optimal Mixed Diagnostic Test With Reference to the Problems of Evolutionary Computation // Proceedings of the First International Conference on Evolutionary Computation and Its Applications (EVCA'96). Moscow, 1996. P. 292-297.

[Potter, De Jong, 2000] M. Potter, K. De Jong. Cooperative coevolution: An architecture for evolving coadapted subcomponents // Evolutionary Computation, 2000, vol. 8, no. 1, pp. 1-29.

[Yankovskaya, Gedike, 1999] A. Yankovskaya, A. Gedike. Finding of All Shortest Column Coverings of Large Dimension Boolean Matrices// Proceedings of the First International Workshop on Multi-Architecture Low Power Design (MALOPD). – ISBN 5-93576-002-9, Moscow, 1999. – pp. 52-60. (http://www.dice.ucl.ac.be/~anmarie/MALOPD).

[Yankovskaya et al., 2003] A.E. Yankovskaya, A.I. Gedike, R.V. Ametov, A.M. Bleikher. IMSLOG-2002 Software Tool for Supporting Information Technologies of Test Pattern Recognition// Pattern Recognition and Image Analysis, 2003, vol. 13, no. 4, pp. 650-657.

Authors' Information

Anna E. Yankovskaya – Tomsk State University of Architecture and Building, 2, Chitinskaya Str, apt. 28, 634003, Tomsk, Russia; e-mail: yank@tsuab.ru

Yury R. Tsoy – Tomsk Polytechnic University, 84, Sovetskaya Str, 634050, Tomsk, Russia; e-mail: qai@mail.ru