Computer Science Journal of Moldova, vol.**9**, no.3(27), 2001

Genetic algorithms for the synthesis optimization of a set of irredundant diagnostic tests in the intelligent system

Anna E. Yankovskaya Alex M. Bleikher

Abstract

Method of the synthesis optimization of a set of irredundant diagnostic tests with genetic algorithms used to solve problems of large dimension is suggested. The idea of creating the irredundant partial implication matrix sectionalized by classification mechanisms lies in the base; creating a set of irredundant diagnostic tests is based on revealing certain kinds of regularities with the use of genetic transformations. All the obligatory and non-informative features are not used in genetic transformations. Procedures of selection able-to-compete individuals from populations, decision making concerning the object under examination of each able-to-compete individual from populations, and organizing of voting on the set of these individuals are suggested. In order to solve these problems the intelligent system is used.

1 Introduction

In this paper we suggest a new approach to solve the problems of pattern recognition by the synthesis optimization of a set of irredundant diagnostic tests (IDT) of a large dimension. The optimization is based on test methods combined with genetic algorithms (GA) and evolutionary computation (EC). A lot of researches devoted to problems of pattern recognition were suggested in the paper (Zhuravlev, Gurevich, 1990). Several original interpretations and versions of GA used in theory of evolutionary computation, were suggested in the papers (Holland, 1975; Goodman, Punch, 1995).

Genetic algorithms began to be used in pattern recognition due to extremely rapid recent development works in the field of soft computations. The term "soft computations", presented by Zadeh (Zadeh, 1994), that includes genetic algorithms, appeared earlier than scientific trend "soft computations" derived from Holland paper (Holland, 1975), where the canonical genetic algorithm was described. There are reviews of papers on genetic algorithms and their development (Skurihin, 1995; Kureichik, 1995) and application of genetic principles in pattern recognition (Shmerko and others, 1995).

The researches for the synthesis optimization of a set of IDT with the use of genetic algorithms are new approaches to evolutionary computation. The first results were published in the paper (Yankovskaya, 1996). In the article (Yankovskaya, 1996) the design of optimal mixed diagnostic tests with the use of genetic transformations on a set of pseudoalternative features (variables) got by means of logic-combinatorial procedures is proposed. Herewith the problem of a test synthesis reduces to the search of all the minimum (optimal) shortest column coverings of Boolean implication matrix, got from the ternary matrix of object description (Q) in the feature space and the integer discrimination matrix R, defining different mechanisms of object partition into classes of equivalence. Optimization and genetic transformations are performed by using cost (fitness) function. The optimization and genetic transformations of implication matrix have been based on using the crossover to a set of features (genes) not entering into a number of constant, obligatory and non-informative features and the procedure of competitive selection of individuals belonging to populations (submatrices of matrix Q). In the paper (Yankovskaya, 1999(1)) the algorithm of unconditional IDT synthesis on the basis of the irredundant matrix implication with the use of genetic algorithms is suggested. In addition, the synthesis of IDT without the construction of the irredundant matrix implication, but on the basis of a step-by-step algorithm is suggested. The step-by-step algorithm is analogous to the algorithm of coding internal states of asynchronous automata (Yankovskaya, 1970).

The problem of pattern recognition by the construction of the shortest column coverings, described in the above mentioned publications,

was solved successfully under a small feature space, where a number of features do not exceed 200-400. With increasing the dimension of a space of features the problems with the computer resource lack (volume of memory and large time required) appeared. These problems caused the creation of an algorithm based on the use of construction of only a part of the implication matrix on the basis of step-by-step algorithm described in the paper (Yankovskaya, 1982). The step-by-step algorithm ensures the reducing of calculations with the use of genetic algorithms.

The synthesis of IDT combined with genetic transformations is performed by means of: the search of irredundant diagnostic tests based on the creation of irredundant partial implication matrix sectionalized on classification mechanisms; the revealing regularities in knowledge and finding the shortest coverings of the matrix; the formation of descendant populations, that will inherit all the obligatory and a part of the informative features of the parent population. And the selection of the samples (individuals) possessing desired features from the previous populations is performed taking into account the criteria of minimization of the gene (feature) number coded by single meanings included into chromosome (test) and maximization of the test weight, which was calculated on the basis of weight feature coefficients and is the cost function. In this paper a brief description of the intelligent system (GenPro) applied for optimizing of synthesis IDT is suggested.

2 Principle notions, knowledge representation and formation of descendant populations

To understand a further interpretation the main notions from publications (Yankovskaya, 1994-1999) are presented.

Nontraditional representation of knowledge in the intelligent system is used: a ternary matrix of description Q and an integer matrix of discrimination R.

The rows of matrix Q are corresponded with the object descriptions and columns are matched with characteristic features. The element q_{ij}

is equal to unity (zero) if the *i-th* object has (has not) the *j-th* feature and q_{ij} is equal to uncertainty (-), where "-" is a symbol of uncertainty treated as a unit and zero values of the given object.

The rows of matrix R are associated with those of the matrix Q under the same name, and columns are associated with classification features that define various mechanisms of classification (mechanisms of partition of objects into classes of equivalency). The element of R in the intersection of the *i-th* row and the *j-th* column defines the membership of the *i-th* object in one of the extracted classes under the *j-th* classification mechanism. The fact that an object belongs to a class is marked by a code number of this class. The rows of Q associated with the same rows of R define the pattern. With a unique classification mechanism, R degenerates into a column, which complies with the traditional representation of information in pattern recognition problems. Notice that a row of Q is associated with a conjunction, and that the element of corresponding row of R defines the number of class where this conjunction assumes a unit value at the corresponding classification mechanism.

This model gives us possibility to present not only data but also expert knowledge, since using only one matrix Q row it is possible to define in an interval form a subset of objects that can be characterized by the same solutions defined by matrix R row.

In crossing pattern descriptions it is found out that matrix Q has to be defined additionally.

Regularities are defined as feature subsets with certain easily interpreted properties, which influence on discrimination of the objects from different patterns, and are stably observed for the objects from the tutorial sample. And at the same time the feature subsets appear on the other objects of the same character. Besides, one of the regularities is the weight coefficient of the features, which characterize their individual contribution in the object discrimination. The subsets being mentioned above include constant, steady (constant within the class), non-informative (differing no pair of objects), alternative (in the meaning of including in a diagnostic test), dependent (in the meaning of including subsets of discriminative object pairs), unessential (included

into no one of the irredundant test), obligatory (included into all of the irredundant tests) features, and also all of the minimal discriminative subsets, which are minimal unconditional irredundant diagnostic tests in their essence.

To present conditions of discrimination of tutorial objects we will use the sectionalized by classification mechanisms binary implication matrix U, constructed on the basis of Q and R matrices. Columns of this matrix U are associated with the characteristic features, and rows are associated with the results of the comparison of different pairs of the objects, which are included into different classes on every classification mechanism. The row of matrix U is the binary vector-function distinguishing. The *m-th* element of matrix U row is equal to 1 (discriminative feature), if the m-th feature, for describing the pairs of objects takes the opposite value (0(1) – in description of the first object, 1(0) – in description of the second object); otherwise it is equal to zero.

To reduce the resources (volume of memory and computations) matrix U is not constructed completely and the search of all the minimal diagnostic tests is reduced to the consecutive construction (with the simultaneous deleting the covering rows) of the irredundant implication matrix U' and finding all the shortest column coverings.

The shortcut construction of the matrix U' is based on the ordering matrix R columns in non-decreasing number of the object pairs to be discriminated by mechanisms of the classification, and on the particular order of choosing the object pairs to be compared, ordered (within classes by means of every mechanism of the classification) by non-decreasing of the number of units in their descriptions. Object pairs from different classes are compared in turns (taken from the beginning and the end of class descriptions) that leads to decrease (without losing regularities) of the number of rows included in the matrix U', that will become covering later, during construction of the matrix. This is cutting the number of the sorting within the comparison (concerning the covering) of each new row in the already constructed part of the matrix U'. Herewith weight coefficients of all the characteristic features are calculated, and at the last step of the construction of

U' matrix, constant, non-informative, pseudoalternative and pseudodependant features are selected. Note that in the earlier papers the term "alternative and dependant features" was used.

The matrix implication will be named irredundant and will be denoted by U', if it has no covering rows (Yankovskaya, 1999(1)).

The matrix implication will be named partial and will be denoted by U'_r, if it represents only some conditions of the matrix implication.

In order to optimize a computer memory the matrix U' is constructed together with the calculations of all weight coefficients of the features. Under the limited resources of memory the matrix U'_r is constructed together with the calculations of all weight coefficients of the features and with the consequent construction of the binary vector-function of distinguishing and with separating the obligatory and constant features.

The weight coefficient w_m of the *m-th* feature corresponding the *m-th* matrix Q column ($m = 1, .., M$) is calculated by the following formula:

$$w_m = \frac{\sum\limits_{r=1}^{K-1} \sum\limits_{t=r+1}^{K} \sum\limits_{i=1}^{N_r} \sum\limits_{j=1}^{N_t} \delta_{ij}^m}{\sum\limits_{i=1}^{K-1} \sum\limits_{j=i+1}^{K} \sigma_i \sigma_j} \tag{1}$$

where K is the number of extracted patterns;

N_f is the number of rows in the *f-th* pattern description ($f \in r, t, i, j$);

$\delta_{ij}^m = 0$, if $q_{im} = q_{jm} = 0$ or $q_{im} = q_{jm} = 1$ (q_{im} is the meaning of the Q matrix element in the cross of the *i-th* row and *j-th* column);

$\delta_{ij}^m = p_i p_j 2^{d_i^- + d_j^-}$, (d_i is the number of values "-" in q_i (row i of the matrix Q), p_i is the replication coefficient *i-th* row), if $q_{im} = 0$ and $q_{jm} = 1$ or $q_{im} = 0$ and $q_{jm} = 0$;

$\delta_{ij}^m = p_i p_j 2^{d_i^- + d_j^- - 1}$, if $q_{im} =$"-" and (or) $q_{jm} =$"-";

σ_j is the number of objects in *j-th* pattern ($j = 1, .., K$), calculated by formula: $\sigma_j = \sum\limits_{l=1}^{N_j} p_l 2^{d_l^-}$.

If $w_m = 0$ then the *m-th* feature is constant.

We will consider the *m-th* feature non-informative, if $w_m < c$, $(m = 1, ..., M)$, where c is the constant, defined from matrices Q and R experimentally.

The search of all the minimal diagnostic tests is reduced to finding the shortest column coverings of the matrix U' by means of the construction of the hierarchic system of the matrix U' submatrices, that is the tree of the search and the shortcut search of all the nonrecurring shortest ways in the search tree, based on the particular labelling of the nodes, that allows to decrease number of sorting and has the essence in the fact, that every level of the search tree contains nodes, corresponding to empty, equal, and used in separation of submatrices columns. Taking into account the first allows to cut branches, that don't correspond with column coverings; the second is to find column coverings without separation of submatrices; the third is to eliminate the construction of the equal the same column coverings.

Irreduntant tests are constructed with genetic transformations use, including obligatory features in all tests and not including constant and non-informative features. The approaches of the construction are described below.

3 Genetic algorithms in test pattern recognition

The function of weight (cost) of a test depends on both the length of the test and weight coefficients of features included in it. The evaluation of minimal length of the test is

$$L =]\log_2 K] \tag{2}$$

where K is the number of extracted patterns;

$]a[$ is the least from above integer to a.

Weight (cost) W_i of the test is feature weight coefficient sum.

A sample (an individual) of the population, that is represented by the submatrix Q with the columns, corresponding to the features,

included into the irredundant diagnostic test, is considered as able-to-compete under the condition of the less number of unit genes (features), included into a chromosome (test), and larger sum of weight coefficients of unit genes, included into this chromosome.

The value of competitive ability depends on the particularities of matrices Q, R. An able-to-compete individual is considered suitable to join the population.

The size (power) of the population is determined by the number of the individuals to be included.

We consider the population more perspective if the ratio of the unit gene weight coefficient sum, included into the chromosome union of the individuals, to the power (number) of the unit genes from this union is more.

The formation of descendant population inheriting features from the previous parent population is founded on the applying of the crossover procedure of chromosomes in a set of genes, matched with non-obligatory features included into IDT (the chromosomes of the parent population). If necessary, each individual may be added by some of features (genes), which are not obligatory or have been included in this individual, from a set of feature union of the parent population. It happens when the result may be not IDT.

Besides, the test is checked up on irredundance. If the test is not irredundant then the redundant feature is deleted. After this operation the individual is included into the population. Adding and deleting features (changing the value of gene of chromosome) realize the mutation operation, which is not applied in the algorithm optimization of IDT synthesis. The size of descendant population depends on the number of genes used in the crossover and on the matrices Q and R.

4 Algorithm of The Synthesis Optimization of a Set of Irredundant Diagnostic Tests

Algorithm of the synthesis optimization of IDT set consists in the following:

1 When computer resources are available: executing the procedure of the construction of the irredundant matrix implication U' and the realization of the algorithm given in the publication (Yankovskaya, 1999(1)) are performed. Otherwise, the construction of only a part of U' (equal U'_r), the calculation of w_m and finding the obligatory features are performed.

1.1. The calculation of the vector-function of distinguishing on the basis of the matrices Q and R. Construction of U' with simultaneous calculation w_m, separation of the obligatory features and deleting the covering rows is performed. If the matrix U' has been constructed completely then the realization of the algorithm described in the publication (Yankovskaya, 1999(1)) is performed.

1.2. Saving the matrix U'_r and fixing the pointer t that is a point of the processed distinguishing conditions from the matrix R.

1.3. Consequent construction of the vector-function of distinguishing from the t-point with a simultaneous calculation of w_m and finding the obligatory features and adding them to the earlier discovered.

2 Construction of the parent population.

2.1. Construction of every individual (submatrix of matrix Q) with the step-by-step algorithm use. The genes are included into chromosome if they are matched by the obligatory features. Also, in order to build IDT, the genes (features which have been generated with a generator of random codes, taking into account weight coefficients w_m) are added to chromosome. In this case the matrix U'_r is used for speeding up construction.

2.2. Selection and including the competitive individuals being matched to all minimal and irredundant tests with maximal

values of W_i into the parent population. The size of the parent population is defined by the experiment. The end of the construction procedure of the parent population depends on the computation time limit or reiteration of already generated individuals.

3 Construction of the next perspective population based on the parent population with the use of the crossover operation and checking up the irredundance of diagnostic tests which match the chromosomes included into the perspective population. The end of the generation of the population is defined by either a building of all IDT or a defined number of populations or the end of time limit.

5 Brief description of the intelligent system

The intelligent system GenPro for the synthesis optimization of IDT with the use of genetic algorithms is realized in Borland C++ Builder language on platform Windows95/NT.

The intelligent system consists of the following components:

1. Initialization.
2. Processing.
3. Representation results.

The initialization involves a process of the construction of knowledge base and of the parameters for the control of generation populations. In particular, the computation time limit and the maximal number of allowed populations are calculated.

To describe the knowledge base the class *PRows* is applied. This class assigns two matrix Q and R. It consists of:

Name is the name of an object (row);

Item is the matrix Q row that is equal to a ternary array. Every item of the array belongs to the set {0,1,2}, where 0 is equal to "0" in the matrix Q, 1 is equal to "1" of the matrix Q and 2 is equal to "-" of the matrix Q;

Count is the number of columns (features) of the matrix Q;

RItem is the matrix R row. Similarly *Item*.

RCount is the number of columns of the matrix R;

NImage is the number of the selected patterns.

Every new object included in the test is described in the system as a class *PRows*. This object is added into the variable *FQobj* that is the pointer of the class TList from VCL (Visual Component Library). *FQobj* saves the matrices Q and R data in program code.

Processing was realized in the system as a class *TGenClass*. It provides the procedures for performing the following operations: the calculation of weight coefficients (*Computewm*), the generation of populations (*Generation*), the selection of competitive individuals (*GetBestInd*) and the check of irredundance (*IsIrredundant*).

It is necessary to note that the procedure of weight coefficients calculation passes within one iteration and involves both the procedure of irredundant implication matrix construction (or partial implication matrix), and the procedure of revealing obligatory and constant features.

The unit of result representation consists of the procedures of saving and printing of the carried out results.

6 Conclusions

The suggested approach realized in the intelligent system provides the synthesis optimization of IDT at large dimension for pattern recognition problems.

The optimization is reached by the construction of the part of irredundant implication matrix which is used for generation of competitive individuals and populations. It is performed within one iteration and

involves the calculation of the weight coefficients of genes, revealing obligatory and constant genes and applying the genetic transformations of chromosomes.

There is a good reason to believe that the method of the synthesis optimization of IDT with the use of genetic algorithms realized in the intelligent system will be widely distributed because it is oriented for a large dimension problems and its parameters depend on the problem dimension. Besides, the purposeful searching is used for competitive individuals and populations.

The presence of the intelligent system realized on Windows platforms and based on the test pattern recognition allows to complement it with new procedures describing other algorithms, with low expenses. The intelligent system allows to test the knowledge bases from different problem areas (medicine (Yankovskaya, 1994), medicine of emergency, geology, ecology, genetics and others).

This work was supported by the Russian Foundation for Basic Research, projects no. 01-01-00772, 01-01-01050.

References

[1] Yu. I. Zhuravlev, I. B. Gurevich (1990). *Pattern Recognition and Image Analysis*. In: Pospelov D.A. (ed.), Artificial Intelligence. Models and Methods, Vol. 2, pp. 149–190. Russia, Moscow: Radio and Connection.

[2] E. D. Goodman, W. F. Punch (1995). *New Techniques to Improve Coarse-grain Parallel GA Performance*, Proceedings of the XXII International Conference "CAD-95: New Information Technologies for Science, Education, Medicine and Business", part 2, pp. 7–15. Ukraine, Yalta (Gurzuf): Crimea.

[3] J. H. Holland (1975). *Adaptation in Natural and Artificial Systems*, University Michigan Press. Ann Arbor.

[4] L. A. Zadeh (1994). *Fuzzy logic, neral network and soft computing*, Communication of the A.C.M,. Vol. 37, **3**: pp. 77–84.

[5] A. M. Skurihin (1995). *Genetic Algorithms*, News of Artificial Intelligence, **4:** pp. 6–46.

[6] V. M. Kureichik (1999). *Genetic Algorithms. Condition. Problems. Perspectives*, Theory and Control systems, **1**, pp. 144–160. Russia, Moscow.

[7] V. Shmerko, S. Yanushkevich, E. Zaitseva (1995). *Genetic Principles in Pattern Recognition and Image Processing*, Proceedings. of the Third International Conf. "Pattern Recognition and Information Analysis". Vol.3: pp.17–24. Belarus, Minsk: SZCZECIN,.

[8] A. E. Yankovskaya (1996). *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", pp. 292–297. Russia, Moscow.

[9] A. E. Yankovskaya (1999(1)). *The Test Pattern Recognition with Genetic Algorithm Use.* In:B. Radig, H. Niemann, Y. Zhuravlev, I. Gourevitch, I. Laptev (Eds.).,5^{th} *Open German-Russian Workshop on Pattern Recognition and Image Understanding,* pp.47–54. Germany, Herrshing.

[10] A. E. Yankovskaya (1970). *Algorithms of Internal States Coding of Asynchronous Automata*, Collection, Digital Models and Integrated Structures, pp. 390–399. Russia, Taganrog.

[11] A. E. Yankovskaya (1982). *Algorithms of Descending for Discrete Field Synthesis Problems and Applications*, Collection, Theory of Discrete Control Devices, pp. 206–214. Russia, Moscow.

[12] A. E. Yankovskaya (1994). *Test Pattern Recognition Medicine Systems with Cognitive Graphics Use*, Computer chronicles, **8/9**: pp. 61–83.

[13] A. E. Yankovskaya (1999(2)). *The Test Pattern Recognition with Genetic Algorithm Use*, Pattern Recognition and Image Analysis, Vol.9, **1:** pp. 121–123.

A.E. Yankovskaya, Alex M. Bleikher, Received November 23, 1999

Anna E. Yankovskaya
Tomsk State University of Architecture and Building,
2 Solyanaya square, 634003,
Tomsk, Russia
E–mail: *yank@tisi.tomsk.su*
Phone: +7(3822) 656924

Alex M. Bleikher
Tomsk Polytechnic University,
30, Lenin avenue, 634034,
Tomsk, Russia
E–mail: *bleikher@chat.ru*
Phone: +7(3822) 418227