УДК 007.52; 519.68:159.955

А.Е. Янковская, Ю.Р. Цой

ПРИМЕНЕНИЕ ГЕНЕТИЧЕСКИХ АЛГОРИТМОВ В ИНТЕЛЛЕКТУАЛЬНЫХ РАСПОЗНАЮЩИХ СИСТЕМАХ[1]

Рассматривается проблема выбора оптимального подмножества безызбыточных безусловных диагностических тестов с использованием генетического алгоритма. Представленные результаты экспериментов для псевдослучайных матриц диагностических тестов показывают высокую сходимость и эффективность предлагаемого подхода.

Ключевые слова: *искусственный интеллект*, *тестовое распознавание образов*, *генетический алгоритм*, *оптимальное подмножество безызбыточных безусловных диагностических тестов*, *интеллектуальные распознающие системы*, *результаты экспериментов*.

В интеллектуальных системах формирование и выбор «хороших» [1] безусловных безызбыточных диагностических тестов (ББДТ) являются одними из наиболее важных шагов при принятии решений, поскольку от свойств используемых тестов существенно зависит качество получаемых решений. Идея использования генетических алгоритмов (ГА) для построения ББДТ при большом признаковом пространстве предложена в статьях [2 – 4]. Первые алгоритмы построения ББДТ, описанные в [2 – 4], программно реализованы и развиты в плане оптимизации построения в последующих работах Янковской А.Е., Блейхер А.М. и др. [5 – 7].

Однако выбор «хороших» ББДТ не всегда приводит к оптимальному решению, поскольку общее количество признаков в выбранном множестве тестов может быть слишком большим, так же как временные и стоимостные затраты или ущерб (риск) [8], наносимый в результате выявления значений признаков исследуемого объекта, например в медицине. В связи с этим предложено применение ГА для построения ББДТ, а также и для формирования субоптимального относительно выбранных критериев подмножества ББДТ.

1. Определения и обозначения

Воспользуемся определениями и обозначениями, необходимыми для постановки задачи и при дальнейшем изложении [3, 4].

Тестом называется совокупность признаков, различающих любые пары объектов, принадлежащих разным образам (классам). Тест называется *безызбыточным*, если при удалении любого признака тест перестает быть таковым. Признак называется *обязательным*, если он содержится во всех безызбыточных тестах. Признак называется *псевдообязательным*, если он не является обязательным и входит во множество используемых при принятии решений безызбыточных тестов.

Пусть $\mathbf{T} = \{t_{ij} : i = 1,...,n, j = 1,...,m\}$ – матрица ББДТ, n – количество ББДТ, m – количество характеристических признаков, булевой строкой $\mathbf{T}_i$ представлен i-й

[1] Работа выполнена при поддержке РФФИ (проект № 07-01-00452) и РГНФ (проект № 06-06-12603В).

ББДТ. Тем же символом $\mathbf{T}_i$ будем обозначать подмножество характеристических признаков, вошедших в ББДТ. Обозначим через $\mathbf{z} = \{z_j : j = 1,...,m\}$ – множество характеристических признаков, причем $t_{ij} = 1 \leftrightarrow z_j \in \mathbf{T}_i$, иначе t_{ij} равно нулю. Для каждого признака z_j зададим весовой коэффициент w_j и коэффициенты стоимости w'_j и ущерба (риска) w''_j [8]. Далее будем использовать термины «вес», «стоимость» и «ущерб» признака вместо соответственно «весовой коэффициент признака, характеризующий его разделяющую способность», «коэффициент стоимости признака, определяющий стоимость выявления его значения» и «коэффициент ущерба, причиняемого в результате выявления значения признака».

Определим вес *i*-го теста:

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

Аналогично определяются значения стоимости и ущерба теста.

2. Постановка задачи

Дана матрица тестов **T** с заданными весами, стоимостью и ущербами признаков. Необходимо выделить такую подматрицу $\mathbf{T}_0$, содержащую n_0 строк, чтобы соответствующее ей множество тестов $\mathbf{N}^0$ обеспечивало выполнение следующих критериев в порядке их следования:

1) во множестве $\mathbf{N}^0$ должно содержаться максимальное число псевдообязательных признаков;

2) множество $\mathbf{N}^0$ должно содержать минимальное общее число признаков;

3) множество $\mathbf{N}^0$ должно иметь максимальный суммарный вес;

4) множество $\mathbf{N}^0$ должно иметь наименьшую суммарную стоимость;

5) множестве $\mathbf{N}^0$ должно иметь наименьший суммарный ущерб.

Поскольку критерии могут противоречить друг другу в зависимости от рассматриваемой прикладной задачи, то искомая подматрица может включать субоптимальное подмножество ББДТ. Приоритеты критериев также зависят от рассматриваемой задачи.

3. Генетический алгоритм

Для решения поставленной задачи предлагается использовать ГА, представляющий итерационный вероятностный эвристический алгоритм поиска. Отличительной особенностью ГА является одновременная работа со множеством точек (популяцией) из пространства потенциальных решений. Каждое возможное решение представлено булевой хромосомой (строкой) длины *n*, каждый *i*-й символ которой кодирует включение *i*-го диагностического теста в итоговое подмножество.

Будем вычислять приспособленность *k*-й особи f_k с хромосомой **h** путем оценки качества соответствующей подматрицы $\mathbf{T}_0(\mathbf{h})$ в соответствии с выражением [8]:

$$f_k = \sum_{k=1}^{5} \mathrm{v}_k e_h^{(k)} + 100\left(U(\mathbf{h}) - n_0\right)^2, \quad f \to \min,$$

где v_k – весовой коэффициент *k*-го критерия, соответствующий его значимости; $U(\mathbf{h})$ – количество единичных разрядов в булевой строке **h**; $M(c)$ – количество единичных столбцов в подматрице $\mathbf{T}_0(\mathbf{h})$, $M(d)$ – количество ненулевых столбцов

в подматрице $\mathbf{T}_0(\mathbf{h})$; $e_h^{(k)}$ – функция штрафа за невыполнение k-го критерия:

$$e_h^{(1)} = \frac{m - M(c)}{m}, \quad e_h^{(2)} = \frac{M(d)}{m}, \quad e_h^{(3)} = \frac{S_W(\mathbf{T}) - S_W(\mathbf{T}_0(\mathbf{h}))}{S_W(\mathbf{T})},$$

$$e_h^{(4)} = \frac{S_{W'}(\mathbf{T}_0(\mathbf{h}))}{S_{W'}(\mathbf{T})}, \quad e_h^{(5)} = \frac{S_{W''}(\mathbf{T}_0(\mathbf{h}))}{S_{W''}(\mathbf{T})},$$

где $S_W(\mathbf{\Psi}), S_{W'}(\mathbf{\Psi})$ и $S_{W''}(\mathbf{\Psi})$ – соответственно суммарный вес, стоимость и ущерб по всем тестам множества, соответствующего матрице $\mathbf{\Psi}$ ($\mathbf{\Psi} \in \{\mathbf{T}, \mathbf{T}_0(\mathbf{h})\}$).

Отметим, что выбор значений штрафов зависит от рассматриваемой прикладной задачи.

4. Результаты экспериментов

Исследование особенностей использования ГА для решения поставленной задачи проведено с использованием псевдослучайных матриц тестов размерностями 1000 × 50, 1000 × 100, 1000 × 200, 1000 × 300, 1000 × 400, 1000 × 500, 2000 × 500. Элементы матриц определяются псевдослучайным образом, после чего производится удаление поглощающих строк. Значения весов, стоимостей и ущербов признаков также определяются как псевдослучайные величины, равномерно распределенные в интервале [0; 1]. Мощность n_0 искомого подмножества тестов для всех экспериментов примем равной 300, что соответствует опыту решения задач в ряде проблемных и междисциплинарных областей (медицина, геология, экобиомедицина, экономика, психология и др.).

Отметим, что псевдослучайное заполнение матриц тестов соответствует отсутствию корреляции между характеристическими признаками, что приводит к минимизации числа возможных закономерностей в исходной матрице тестов. В силу этого использование псевдослучайных матриц тестов представляет более сложную по сравнению с реальной задачу.

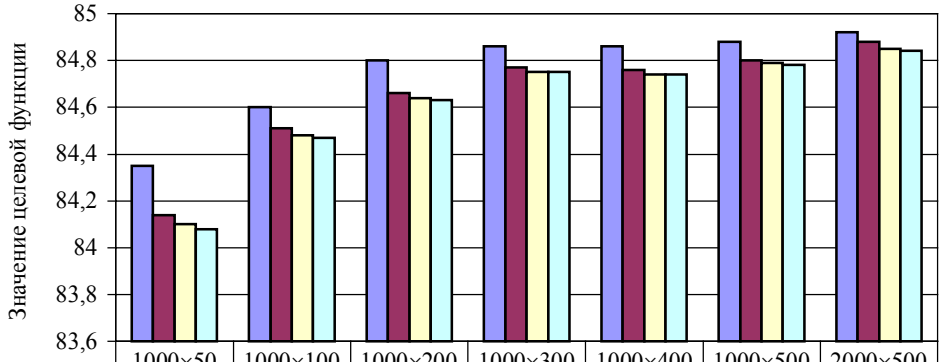
Значения штрафов установлены следующим образом: $v_1 = 40$, $v_2 = 30$, $v_3 = 15$, $v_4 = 10$, $v_5 = 5$. Отметим, что используемые значения штрафов выбраны безотносительно прикладной задачи. Основным критерием их выбора является соответствие приоритету критериев оптимизации, сформулированных выше. Рассматривается ГА с турнирной селекцией при размере турнира равном 6, двухточечным оператором кроссинговера, битовой мутацией и 1 элитной особью. По итогам 100 независимых запусков для каждой из рассматриваемых матриц будем оценивать результаты как по полученному лучшему значению функции приспособленности, так и по следующим критериям, сформулированным в [9] и характеризующим стабильность решений, полученных в различных запусках:

1. Критерию стабильности, учитывающему частоту p_i встречаемости i-го теста во всех решениях, полученных по результатам 100 запусков ГА. Чем больше количество тестов, для которых значение p_i равно или близко к 1, тем выше сходимость алгоритма.

2. Суммарному количеству Ω ББДТ, не вошедших в полученные решения. Чем больше Ω, тем выше сходимость алгоритма.

Представленные критерии стабильности и сходимости будут использоваться для оценки работы ГА, и их введение не продиктовано особенностями используемых матриц ББДТ.

Полученные лучшие значения целевой функции, усредненные по 100 запускам, для различных матриц ББДТ в зависимости от размера популяции показаны на рис. 1. Поскольку рассматривается задача минимизация целевой функции, то можно отметить улучшение результатов при увеличении размера r популяции, однако это улучшение весьма незначительно, в большинстве случаев порядка 10^{-2}.



	1000×50	1000×100	1000×200	1000×300	1000×400	1000×500	2000×500
r = 20	84,35	84,6	84,8	84,86	84,86	84,88	84,92
r = 50	84,14	84,51	84,66	84,77	84,76	84,8	84,88
r = 100	84,1	84,48	84,64	84,75	84,74	84,79	84,85
r = 200	84,08	84,47	84,63	84,75	84,74	84,78	84,84

Размерность матрицы тестов

Рис. 1. Результаты решения поставленной задачи в зависимости от размера популяции для псевдослучайных матриц различной размерности

Отметим, что время работы ГА в зависимости от размера популяции зависит линейно (рис. 2). Исходя из этого, при решении рассматриваемой задачи повышение размера популяции во многих случаях приводит к неоправданному росту вычислительной сложности.

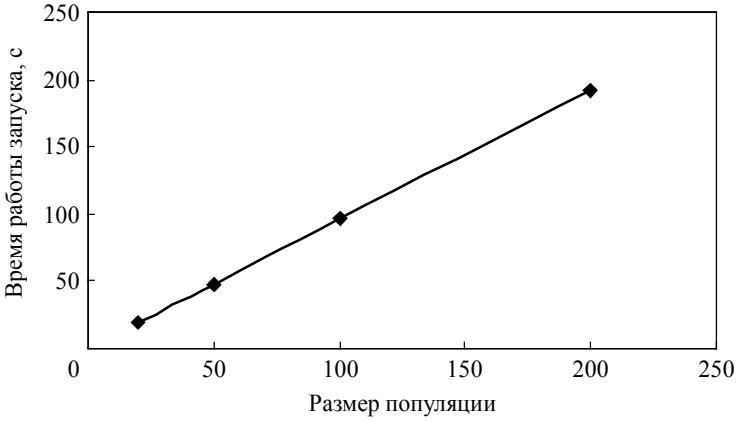
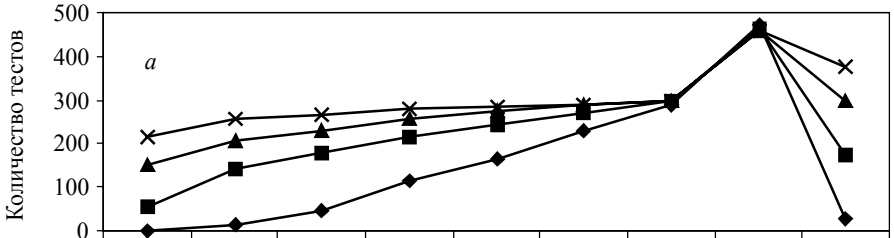


Рис. 2. Зависимость времени работы запуска ГА от размера популяции

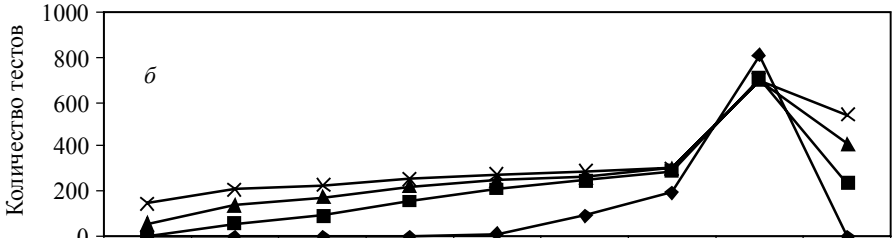
Зависимость количества тестов от частоты их встречаемости для матриц 1000 × 50 и 1000 × 500 в полученных решениях представлена на рис. 3, r – обозначает размер популяции. По оси абсцисс отложен процент встречаемости тестов, а по оси ординат – соответствующее количество тестов. Видно, что с ростом размера популяции сходимость увеличивается, так как растет количество тестов, встречающихся во всех решениях.

Отметим, что в случаях, когда количество тестов, встречающихся в большинстве решений, существенно меньше мощности n_0 искомого подмножества тестов, размер популяции является недостаточным. Примером является случай использования популяции из 20 особей при исходной матрице 1000 × 500, график для которого показан на рис. 3, *б*. Также заметим, что с увеличением количества признаков в исходной матрице тестов сложность задачи увеличивается, что видно из сравнения графиков на рис. 3, *а* и *б*.



	100%	>=95%	>=90%	>=80%	>=70%	>=60%	>=50%	<50%	0%
r = 20	1	16	47	115	164	231	287	473	29
r = 50	56	140	177	216	245	272	296	464	174
r = 100	151	207	228	256	276	289	299	461	298
r = 200	215	259	267	279	284	291	299	461	377

Встречаемость тестов

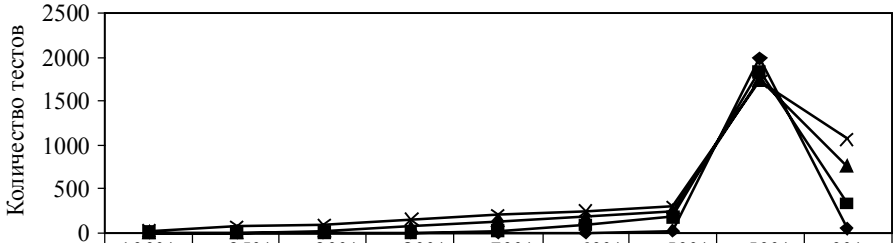


	100%	>=95%	>=90%	>=80%	>=70%	>=60%	>=50%	<50%	0%
r = 20	0	0	0	0	13	92	196	804	0
r = 50	3	56	94	156	207	245	298	702	242
r = 100	55	140	170	218	248	268	300	700	414
r = 200	145	207	233	254	274	290	300	700	538

Встречаемость тестов

Рис. 3. Зависимость количества тестов от частоты их встречаемости в полученных решениях: *а* – результаты для матрицы тестов размерностью 1000 × 50; *б* – результаты для матрицы тестов размерностью 1000 × 500

Зависимости количества тестов от их встречаемости для матрицы ББДТ размерностью 2000 × 500 представлены на рис. 4. Увеличение количества тестов существенно усложняет задачу для ГА, поскольку только для популяции из 200 особей количество тестов со встречаемостью не менее 50 % близко к мощности искомого подмножества тестов.

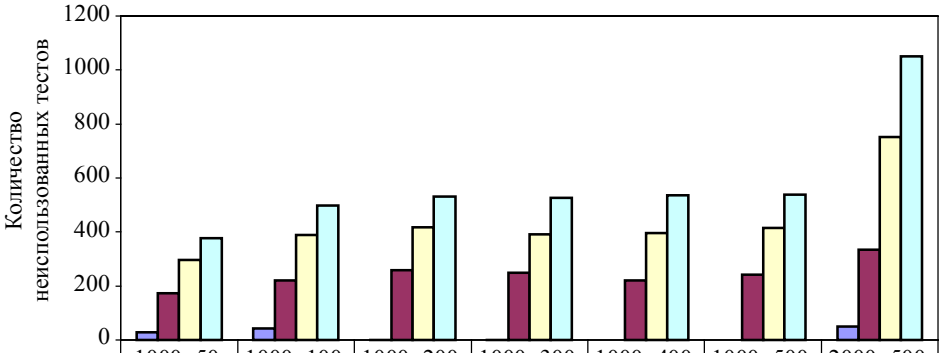


	100%	>=95%	>=90%	>=80%	>=70%	>=60%	>=50%	<50%	0%
r = 20	0	0	0	0	0	2	16	1984	51
r = 50	0	0	0	2	33	86	169	1831	334
r = 100	0	6	23	72	127	179	254	1746	752
r = 200	14	63	90	146	202	238	281	1719	1051

Встречаемость тестов

Рис. 4. Зависимость количества тестов от частоты их встречаемости в полученных решениях для матрицы 2000 × 500

На рис. 5 показана зависимость количества Ω неиспользуемых тестов от размерности матрицы тестов. Также видно, что с ростом размера популяции сходимость работы алгоритма улучшается.



	1000×50	1000×100	1000×200	1000×300	1000×400	1000×500	2000×500
r = 20	29	44	1	1	0	0	51
r = 50	174	220	259	250	221	242	334
r = 100	298	389	417	392	396	414	752
r = 200	377	498	530	526	536	538	1051

Размерность матрицы тестов

Рис. 5. Зависимость количества Ω неиспользованных тестов от размерности матрицы тестов

Анализ решений, полученных при различных настройках ГА, показал, что сформированные по 100 запускам подмножества тестов, соответствующие различным параметрам ГА, отличаются незначительно. Например, для матрицы тестов 1000 ×500 при размерах популяции 50 и 200 особей полученные подмножества тестов отличались только на 35 тестов, что позволяет сделать вывод о достаточно высокой степени сходимости алгоритма. Однако значительное количество тестов, встречающихся менее чем в 50 % решений (соответственно 460 и 162 для популяций из 50 и 200 особей) свидетельствует о возможности повышения эффективности работы ГА и сходимости результатов.

Также было проведено исследование зависимости состава подмножества тестов, сформированного по результатам нескольких запусков ГА, от количества запусков. При использовании матрицы тестов размерностью 1000 × 500 результаты ГА с популяцией размером 50 особей для 10, 20, 30, 40, 50, 60, 70, 80, 90 и 100 запусков совпадают для 245 тестов (из 300 искомых). Совпадение с результатами ГА с популяцией 200 особей составляет 244 теста. Другими словами, 245 и 244 теста присутствуют в большинстве найденных решений, несмотря на различное количество запусков и размер популяции.

Распределение количества тестов в зависимости от частоты их встречаемости для ГА с популяцией 50 особей показано на рис. 6, шкала ординат – логарифмическая. Рост количества тестов, встречающихся во всех решениях, с уменьшением числа запусков можно объяснить усилением роли случайности при малом числе запусков, по которым проводится анализ результатов.

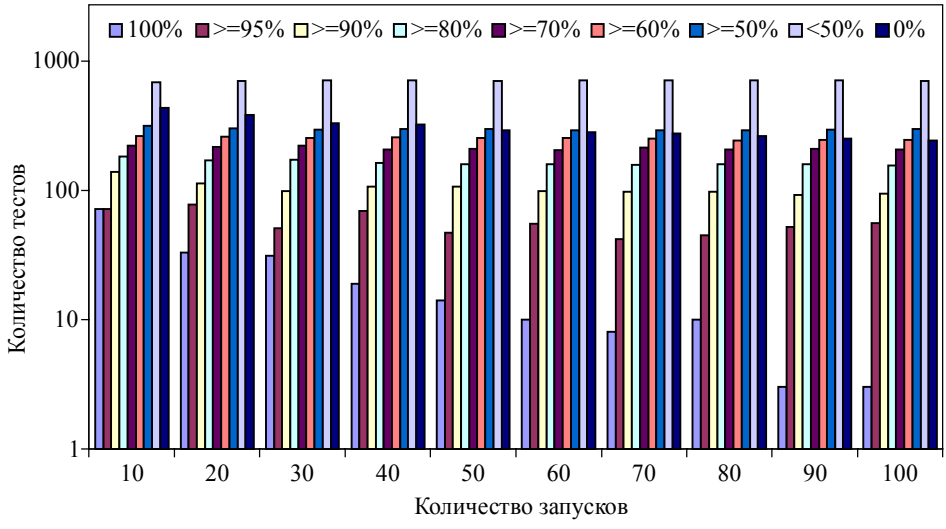


Рис. 6. Распределения количества тестов по частоте их встречаемости в полученных решениях для различного количества запусков ГА для матрицы размерностью 1000 × 500

Таким образом, на основании результатов исследования можно сделать следующий вывод:

Несмотря на то, что увеличение размера популяции способствует повышению сходимости ГА по критериям из работы [9], получены результаты, свидетельст-

вующие о том, что для матриц тестов, имеющих не больше 1000 строк, анализ решений, полученных при использовании сравнительно небольшого размера популяции и малого количества запусков, позволяет сформировать подмножество тестов, близкое к оптимальному.

Данный вывод представляется авторам статьи весьма важным, так как показывает, что возможно эффективное решение поставленной задачи с использованием сравнительно небольших вычислительных затрат. Однако данный вывод необходимо проверить на реальных данных.

В силу приведенного выше анализа результатов сокращение количества особей в популяции в α_1 раз и количества запусков ГА в α_2 раз позволяет уменьшить вычислительные затраты и время поиска решения пропорционально произведению $\alpha_1\alpha_2$.

Заключение

В работе рассматривалось применение ГА для решения задачи формирования субоптимального подмножества ББДТ. Представленные результаты экспериментов показывают достаточно высокую сходимость ГА при решении поставленной задачи.

На основании полученных результатов и их анализа сделан вывод о возможности существенного уменьшения вычислительной сложности ГА при решении рассматриваемой задачи путем уменьшения размера популяции, а также количества запусков. Отметим, что остается неясным вопрос о зависимости минимального допустимого размера популяции и количества запусков от размера и характеристик матрицы тестов, при которых возможно получение решения, близкого к оптимальному.

Дальнейшие исследования будут направлены на разработку более эффективных процедур эволюционного поиска субоптимального подмножества ББДТ для решения задач принятия решений на основе тестового распознавания образов.

ЛИТЕРАТУРА

1. *Naidenova R.A.*, *Plaksin M.V.*, *Shagalov V.L.* Inductive inferring all good classification test // Знание – Диалог – Решение: Сб. науч. тр. Междунар. конф. Ялта, 1995. Т. 1. С. 79 – 84.
2. *Янковскаа А.Е.* Тестовое распознавание образов с использованием генетических алгоритмов // Распознавание образов и анализ изображений: новые информационные технологии (РОАИ-4-98): Труды IV Всероссийской с международным участием конференции. Новосибирск, 1998. Ч. I. С. 195 – 199.
3. *Yankovskaya A.E.* Test pattern recognition with the use of genetic algorithms // Pattern Recognition and Image Analysis. 1999. V. 9. No. 1. P. 121 – 123.
4. *Yankovskaya A.E.* The test pattern recognition with genetic algorithms use // Proc. of the Pattern Recognition and Image Understanding. 5th Open German-Russian Workshop. 1999. P. 47 –54.
5. *Янковская А.Е.*, *Блейхер А.М.* Оптимизация синтеза безызбыточных диагностических тестов с использованием генетических алгоритмов и реализация ее в интеллектуальной системе // Искусственный интеллект. Научно-теоретический журнал. Донецк, 2000. № 2. С. 272 – 278.
6. *Yankovskaya A.E.*, *Bleikher A.M.* Genetic algorithms for the synthesis optimization of a set of irredundant diagnostic tests in the intelligent system // Computer Science Journal of Moldova. 2001. V. 9. No. 3(27). P. 336 – 349.
7. *Yankovskaya A.E.*, *Gedike A.I.*, *Ametov R.V.*, *Bleikher A.M.* IMSLOG-2002 software tool for supporting information technologies of test pattern recognition // Pattern Recognition and Image Analysis. 2003. V. 13. No. 4. P. 650 – 657.

8. *Yankovskaya A.E.*, *Tsoy Y.R.* Optimization of a set of tests selection satisfying the criteria prescribed using compensatory genetic algorithm // Proc. of IEEE EWDTW'05. Kharkov: SPD FL, 2005. P. 123 – 126.
9. *Янковская А.Е.*, *Цой Ю.Р.* Исследование эффективности генетического поиска оптимального подмножества безызбыточных тестов для принятия решений // Искусственный интеллект. Украина, Донецк: ІПШІ «Наука і освіта», 2006. № 2. С. 257 – 260.

Янковская Анна Ефимовна
Томский государственный архитектурно-строительный университет
E-mail: yank@tsuab.ru
Цой Юрий Робертович
Томский политехнический университет
E-mail: qai@mail.ru

Поступила в редакцию 22 января 2009 г.