# Генетический алгоритм и его модификация для формирования оптимального подмножества тестов

Янковская А.Е.
Томский государственный архитектурно-строительный университет
yank@tsuab.ru

Цой Ю.Р.
Томский политехнический университет
qai@mail.ru

# Содержание доклада

1. Введение
2. Постановка задачи
3. Генетический алгоритм
4. Результаты экспериментов
5. Модифицированный алгоритм
6. Заключение

# 1. Введение

Формирование и выбор «хороших» [1] безусловных безызбыточных диагностических тестов (ББДТ) является одним из наиболее важных при принятии решений в интеллектуальных системах, 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, 3, 4]. Первые алгоритмы построения ББДТ, описанные в [2, 3], программно реализованы и развиты в плане оптимизации построения в последующих работах Янковской А.Е. и Янковской А.Е. с Блейхер А.М. [5, 6].

# 1. Введение

Выбор «хороших» безызбыточных диагностических тестов является одним из наиболее важных при принятии решений в интеллектуальных системах. Однако, этот выбор не всегда приводит к оптимальному решению.

Некоторые причины:

- Слишком большое общее количество признаков в тестах.
- Слишком большие затраты (временные, стоимостные) при выявлении значений используемых признаков.
- Слишком большой ущерб (риск) [8], связанный с выявлением значений признаков.

# 2. Основные понятия. Постановка задачи

*Тестом* называется совокупность признаков, различающих любые пары объектов, принадлежащих разным образам (классам).

Тест называется *безызбыточным*, если при удалении любого признака тест перестает быть тестом.

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.

# 2. Основные понятия. Постановка задачи

Пусть $\mathbf{T} = \{t_{ij} : i = 1,...,n,\ j = 1,...,m\}$ – матрица ББДТ, $n$ – количество ББДТ, $m$ – количество характеристических признаков. $\mathbf{T}_i$ соответствует $i$-му ББДТ ($i$-я строка матрицы $\mathbf{T}$). Обозначим через

$\mathbf{z} = \{z_j : j = 1,...,m\}$ – множество характеристических признаков, причем $t_{ij} = 1 \leftrightarrow z_j \in \mathbf{T}_i$. Для каждого признака $z_j$ зададим весовой коэффициент $w_j$ и коэффициенты стоимости $w'_j$ и ущерба (риска) $w''_j$.

# 2. Основные понятия. Постановка задачи

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

Строки матрицы сопоставлены тестам, столбцы – признакам.

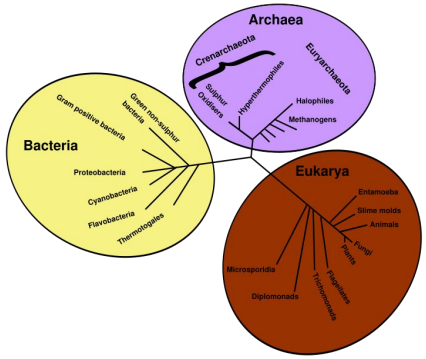
При решении практических задач количество тестов и признаков может принимать значения от нескольких десятков до нескольких десятков тысяч.

# 2. Основные понятия. Постановка задачи

Для данной матрицы $\mathbf{T}$ 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:

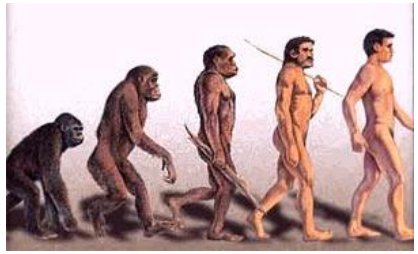
1. В выбранном множестве тестов $\mathbf{N}^0$ мощности $n_0$ должно содержаться максимальное число псевдообязательных признаков.
2. Выбранное множество тестов $\mathbf{N}^0$ должно содержать минимальное общее число признаков.
3. Выбранное множество тестов $\mathbf{N}^0$ должно иметь максимальный суммарный вес.
4. Множество выбранных тестов $\mathbf{N}^0$ должно иметь наименьшую суммарную стоимость.
5. Множество выбранных тестов $\mathbf{N}^0$ должно иметь наименьший суммарный ущерб.

# 3. Генетический алгоритм



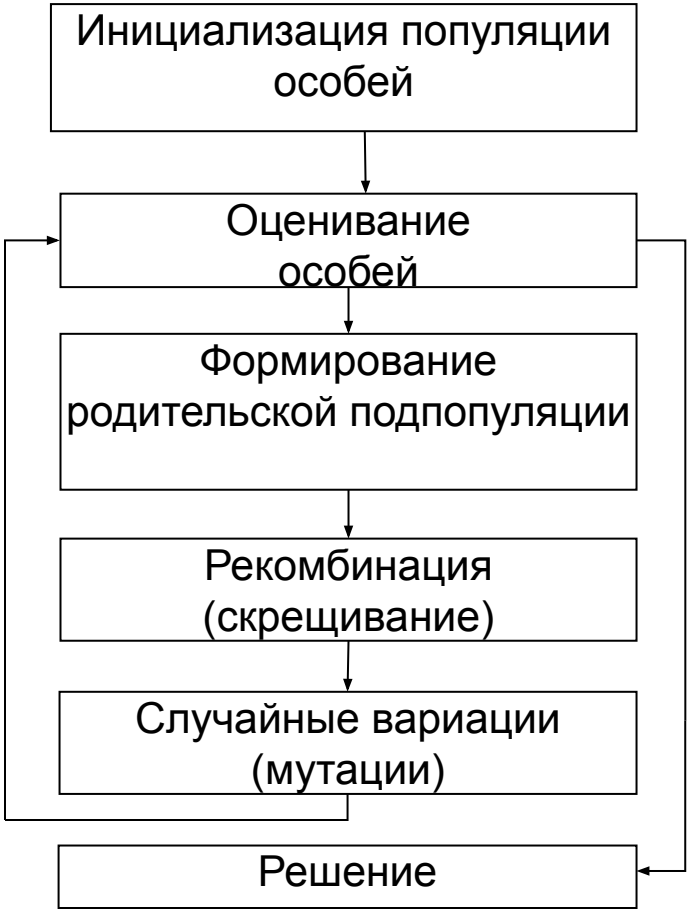

*Эволюционные вычисления (evolutionary computation)* – раздел *мягких вычислений* (soft computing) и *вычислительного интеллекта* (computational intelligence), посвященный разработке методов и алгоритмов, опирающихся на эволюционные принципы наследственности, изменчивости и естественного отбора.

*Генетический алгоритм (ГА) (genetic algorithm)* – вид *эволюционного алгоритма (evolutionary algorithm)*, в котором наибольшее внимание уделяется оператору рекомбинации (скрещивания) возможных решений.



# 3. Генетический алгоритм

Генетический алгоритм




## Терминология

| | |
|---|---|
| *Решение* | *Особь* |
| *Закодированное решение* | *Хромосома* |
| *Множество решений* | *Популяция* |
| *Качество решения* | *Приспособленность особи* |
| *Отбор решений* | |
| *Рекомбинация решений* | *Селекция* |
| *Вариации решений* | *Скрещивание* |
| *Одна итерация (этап)* | *Мутации* |
| | *Поколение* |

# 3. Генетический алгоритм

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

$$
T_0 = \begin{array}{c} \\ 2 \\ 4 \\ 6 \end{array}
\begin{array}{c}
\begin{array}{cccccccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 \end{array} \\
\begin{bmatrix}
0 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \\
0 & 1 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 1 & 0 & 0 \\
1 & 0 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 0
\end{bmatrix}
\end{array}
$$

Исходная матрица **T** и подматрица $\mathbf{T}_0$ (решение)

Закодированное решение: {0,1,0,1,0,1,0}

# 3. Генетический алгоритм

Функция приспособленности:

$$f_k = \sum_{j=1}^{5} v_j e_h^{(j)} + 100\left(U(h) - n_0\right)^2$$

$$f \to \min$$

$v_j$ – весовой коэффициент $j$-го критерия

$U(\boldsymbol{\psi})$ – количество единичных разрядов в бинарной строке $\boldsymbol{\psi}$

$$e_h^{(1)} = \frac{m - U_c\left(\mathbf{T}_0(h)\right)}{m}, \quad e_h^{(2)} = \frac{U_d\left(\mathbf{T}_0(h)\right)}{m}, \quad e_h^{(3)} = \frac{S_W(\mathbf{T}) - S_W(\mathbf{T}_0(h))}{S_W(\mathbf{T})},$$

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

# 3. Генетический алгоритм

$S_W(\mathbf{\Psi}), S_{W'}(\mathbf{\Psi})$ и $S_{W''}(\mathbf{\Psi})$ – соответственно суммарный вес, стоимость и ущерб по всем тестам множества, соответствующего матрице $\mathbf{\Psi}$;

$U_c(\mathbf{\Psi}) = U\left(\bigwedge_i \boldsymbol{\psi}_i\right)$ и $U_d(\mathbf{\Psi}) = U\left(\bigvee_i \boldsymbol{\psi}_i\right)$ – соответственно количество единичных разрядов в конъюнкции и дизъюнкции по всем строкам бинарной матрицы $\mathbf{\Psi}$.

Для экспериментов будем использовать следующие значения весов штрафов: $v_1 = 40$, $v_2 = 30$, $v_3 = 15$, $v_4 = 10$, $v_5 = 5$.

# 3. Генетический алгоритм

Поскольку необходимо максимизировать максимальное количество псевдообязательных признаков в искомом подмножестве ББДТ (критерий 1), а также его суммарный вес (критерий 3), но рассматривается задача минимизации целевой функции *f* , то в выражениях для соответствующих критериям функций штрафа $e_h^{(1)}$ и $e_h^{(3)}$ используется вычитание количества псевдообязательных признаков и веса от максимальных значений. Аналогичные рассуждения использовались при выборе вида функций штрафов для критериев 2, 4 и 5.

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 1000x50, 1000x100, 1000x200, 1000x300, 1000x400, 1000x500, 2000x500. Мощность $n_0$ подмножества ББДТ, которое необходимо сформировать из матрицы $\mathbf{T}$, для всех экспериментов равна 300.

<u>Параметры ГА:</u>

Длительность эволюционного поиска: 1000 поколений

Турнирная селекция, размер турнира равен 6 особям.

Оператор кроссинговера: двухточечный

Оператор мутации: битовый

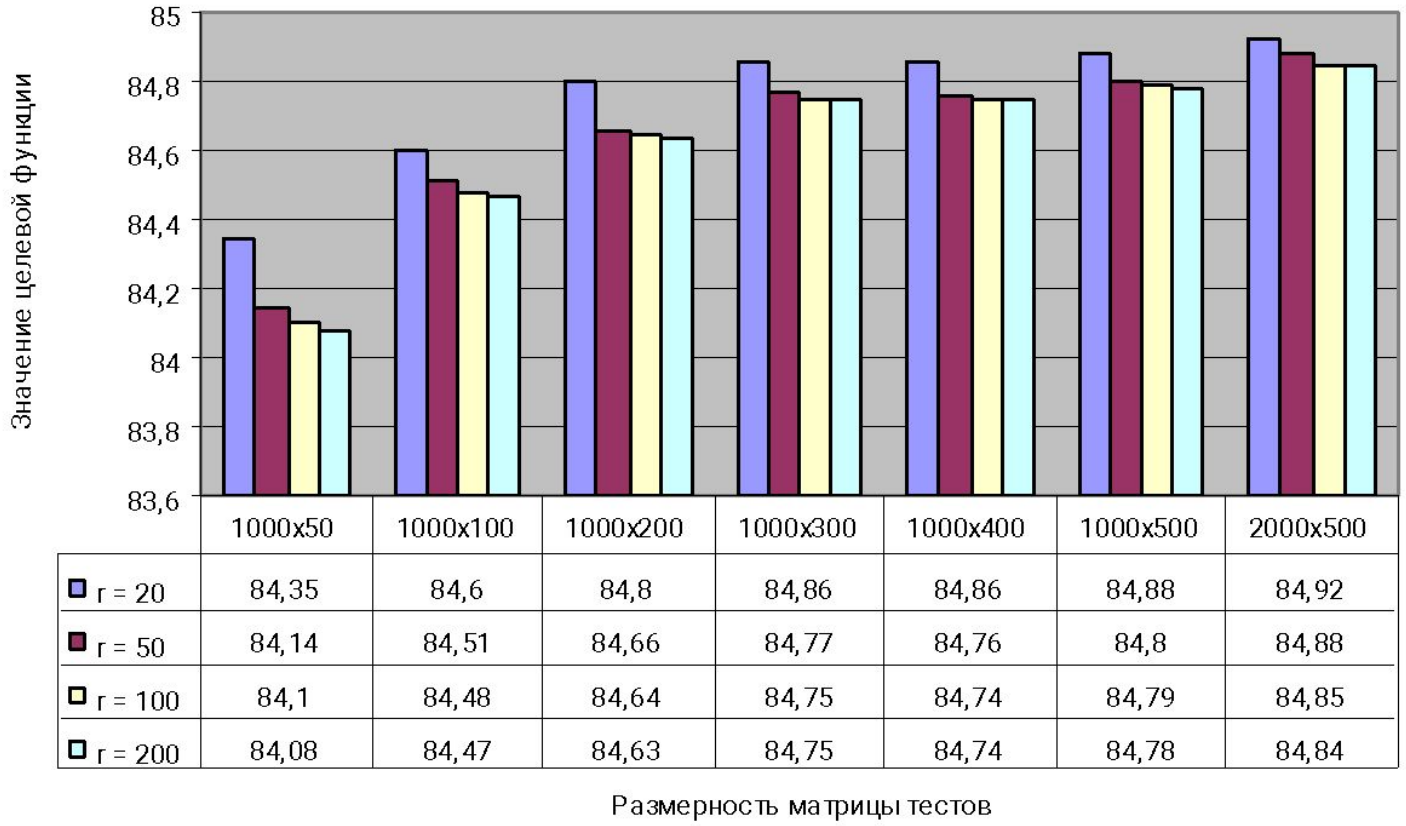
Все результаты усреднены по 100 запускам.

# 4. Результаты экспериментов

Будем оценивать результаты как по полученному лучшему значению функции приспособленности, так и по следующим критериям [11], характеризующим стабильность решений, полученных в различных запусках:

1. Критерий стабильности, учитывающий 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. Суммарное количество Ω ББДТ, не вошедших в полученные решения. Чем больше Ω, тем выше сходимость алгоритма.

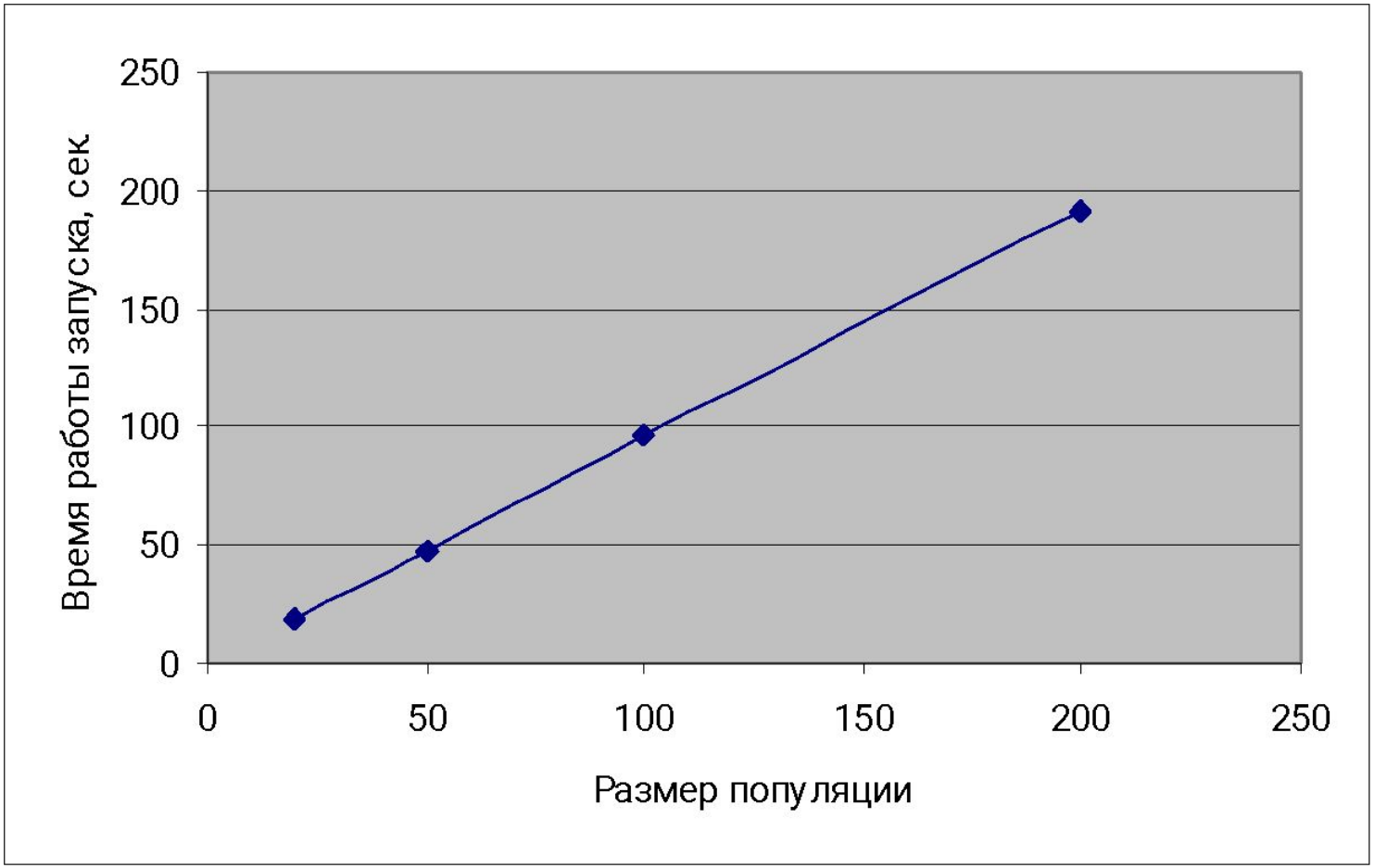
# 4. Результаты экспериментов




| | 1000x50 | 1000x100 | 1000x200 | 1000x300 | 1000x400 | 1000x500 | 2000x500 |
|---|---|---|---|---|---|---|---|
| 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 |

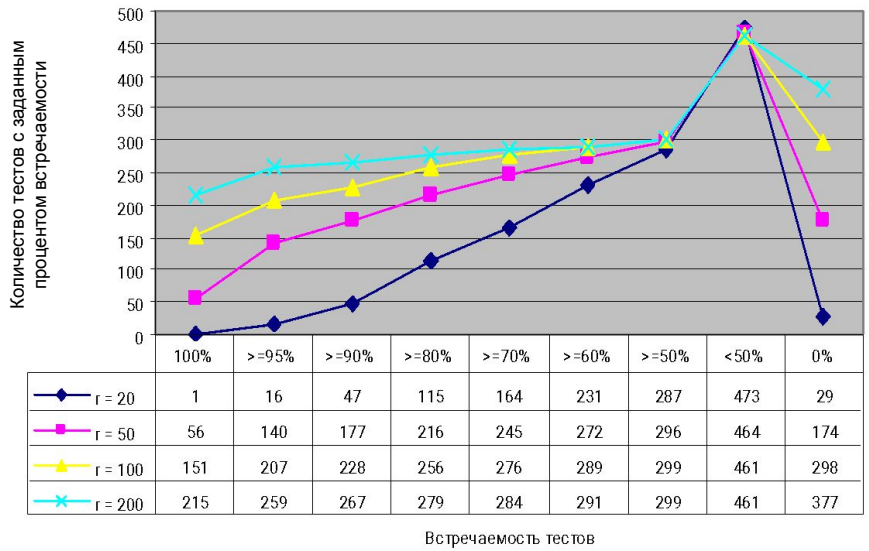
Результаты решения поставленной задачи в зависимости от размера *r* популяции для псевдослучайных матриц различной размерности

# 4. Результаты экспериментов



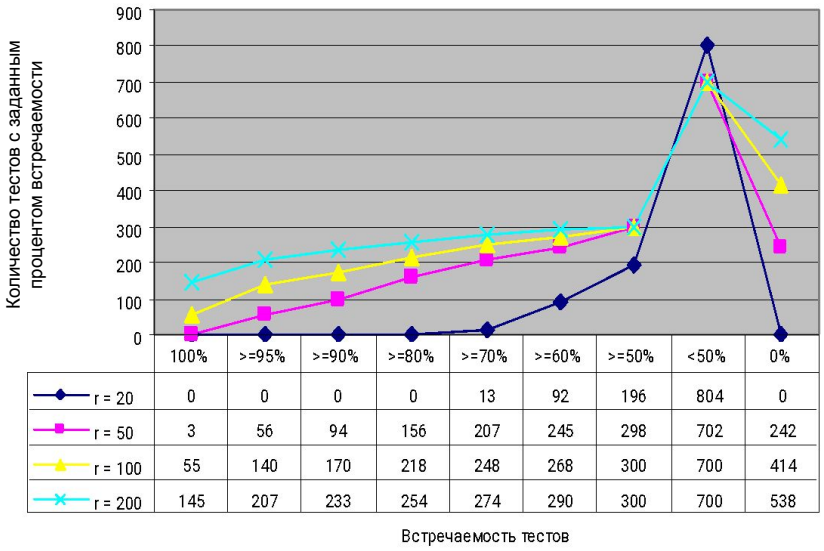

Зависимость времени работы запуска ГА от размера популяции

# 4. Результаты экспериментов




| | 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 |

а) результаты для матрицы тестов размерностью 1000x50

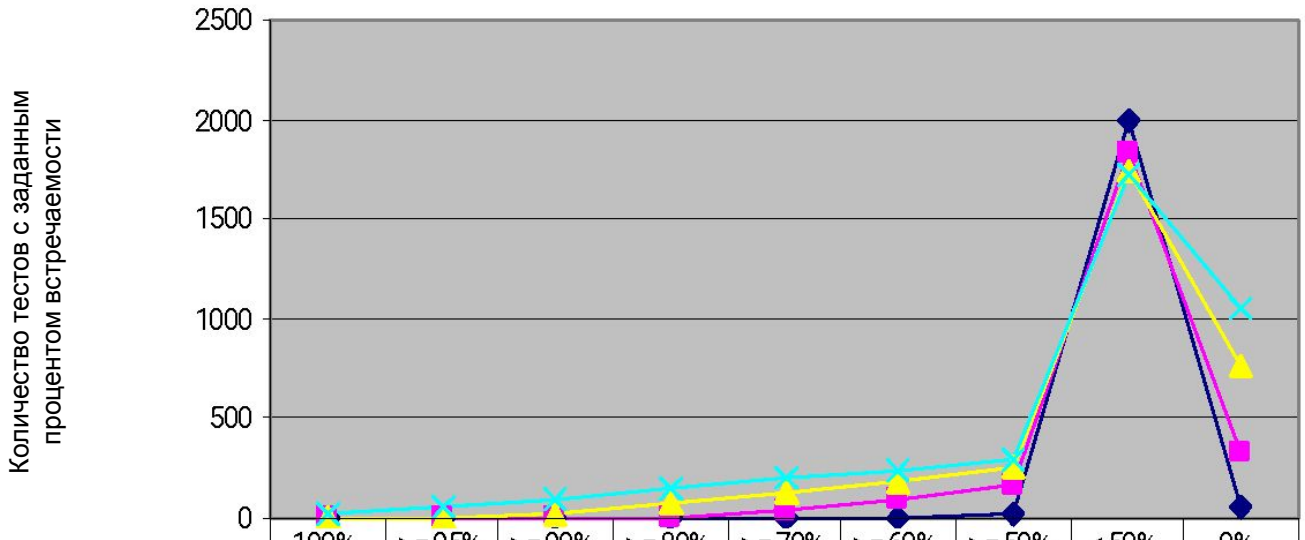



| | 100% | >=95% | >=90% | >=80% | >=70% | >=60% | >=50% | <50% | 0% |
|---|---|---|---|---|---|---|---|---|---|
| r = 20 | 0 | 0 | 0 | 0 | 13 | 92 | 196 | 804 | 0 |
| r = 50 | 3 | 56 | 94 | 156 | 207 | 245 | 298 | 702 | 242 |
| r = 100 | 55 | 140 | 170 | 218 | 248 | 268 | 300 | 700 | 414 |
| r = 200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

б) результаты для матрицы тестов размерностью 1000x500

Зависимость количества тестов от частоты их встречаемости в полученных решениях.
*r* обозначает размер популяции.

# 4. Результаты экспериментов

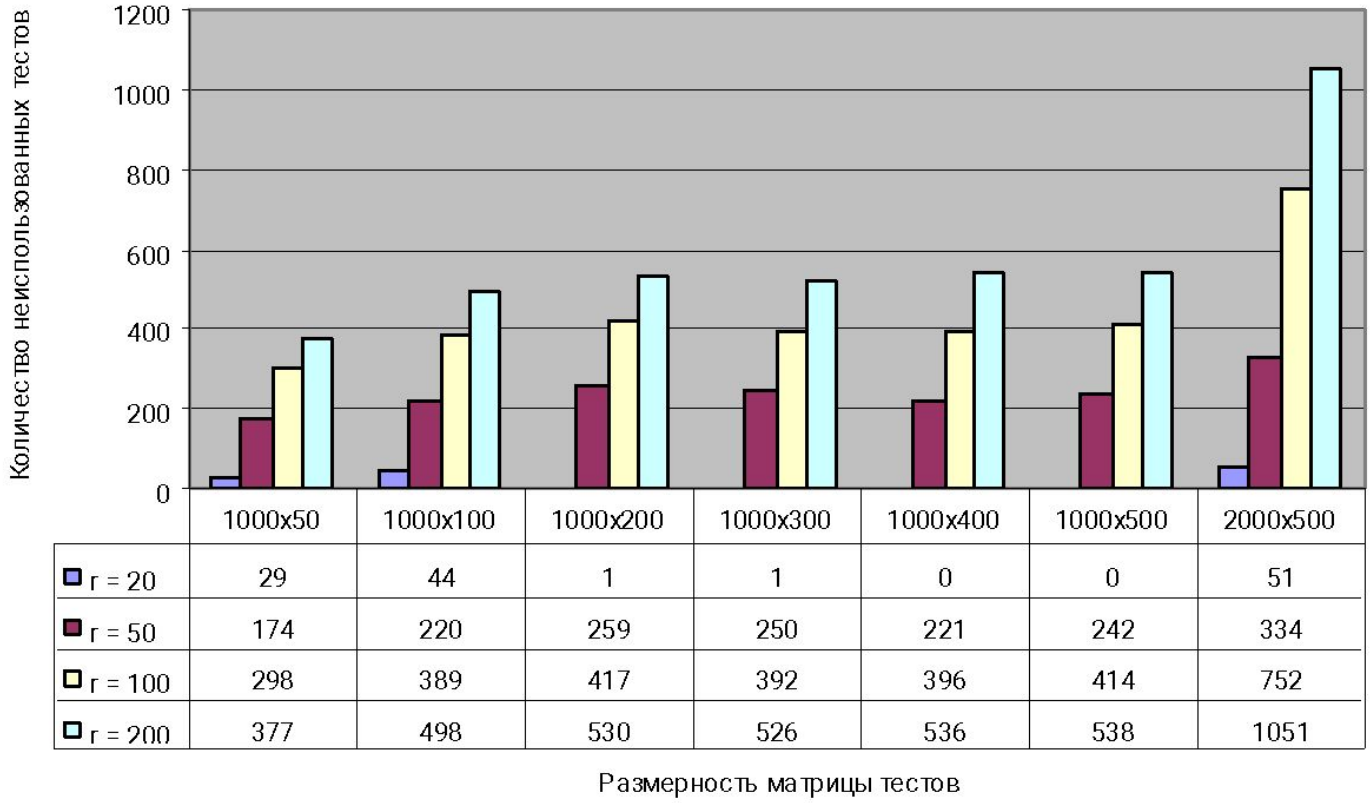



| | 100% | >=95% | >=90% | >=80% | >=70% | >=60% | >=50% | <50% | 0% |
|---|---|---|---|---|---|---|---|---|---|
| r = 20 | 0 | 0 | 0 | 0 | 0 | 2 | 16 | 1984 | 51 |
| r = 50 | 0 | 0 | 0 | 2 | 33 | 86 | 169 | 1831 | 334 |
| r = 100 | 0 | 6 | 23 | 72 | 127 | 179 | 254 | 1746 | 752 |
| r = 200 | 14 | 63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

Встречаемость тестов

Зависимость количества тестов от частоты их встречаемости в полученных решениях для матрицы 2000x500. *r* обозначает размер популяции.

# 4. Результаты экспериментов




| | 1000x50 | 1000x100 | 1000x200 | 1000x300 | 1000x400 | 1000x500 | 2000x500 |
|---|---|---|---|---|---|---|---|
| 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 |

Зависимость количества неиспользованных тестов от размерности матрицы тестов. *r* обозначает размер популяции.

# 4. Результаты экспериментов

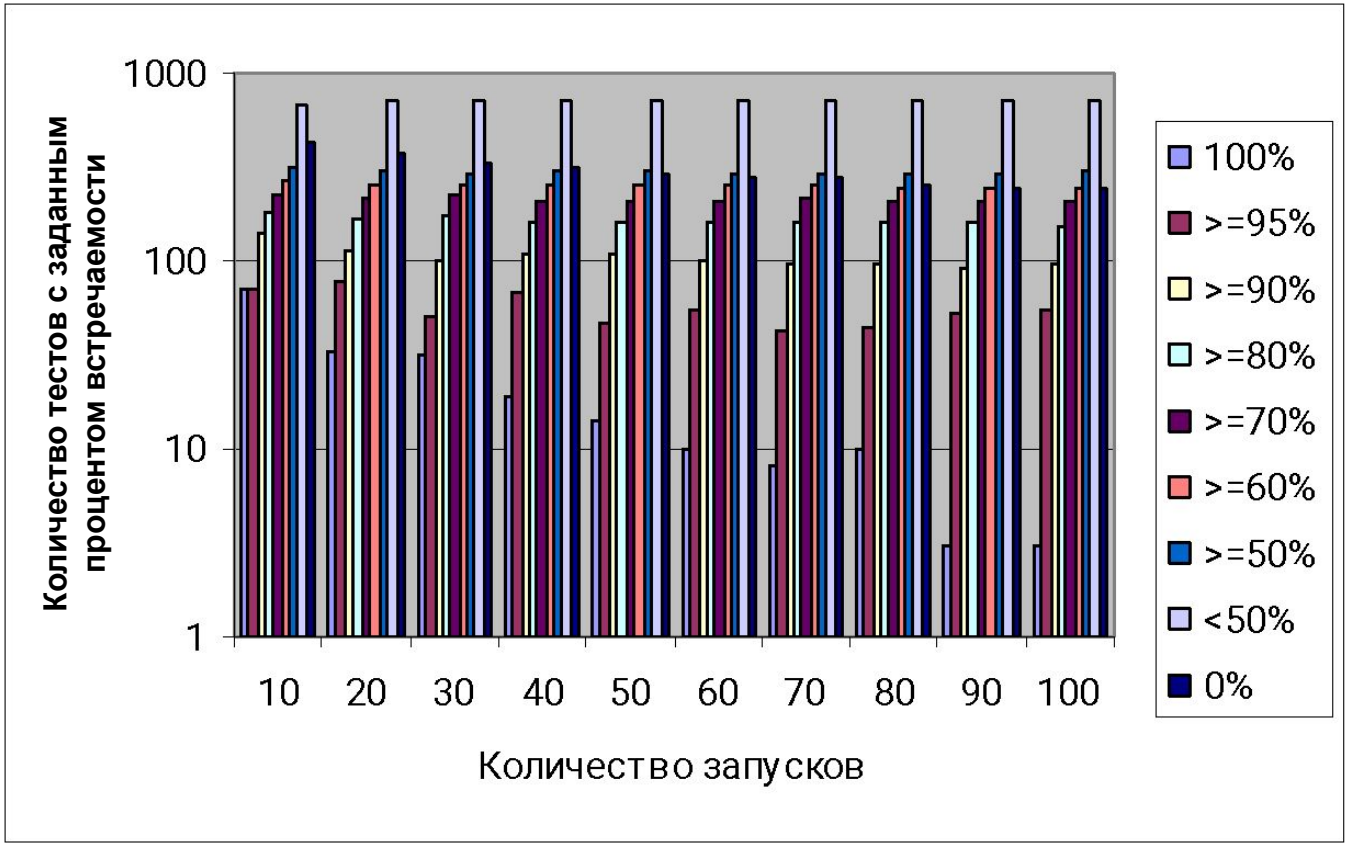
Анализ решений, полученных при различных настройках ГА, показал, что сформированные по 100 запускам подмножества тестов, соответствующие различным параметрам ГА, отличаются незначительно.

Например, для матрицы тестов 1000x500 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 количество тестов, встречающихся менее чем в 50% решений довольно велико (соответственно, 460 и 162 для популяций из 50 и 200 особей).

# 4. Результаты экспериментов

При использовании матрицы тестов размерностью 1000x500 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.

# 4. Результаты экспериментов




Распределения количества тестов по частоте их встречаемости в полученных решениях для различного количества запусков ГА для матрицы размерностью 1000x500. *r* обозначает размер популяции.

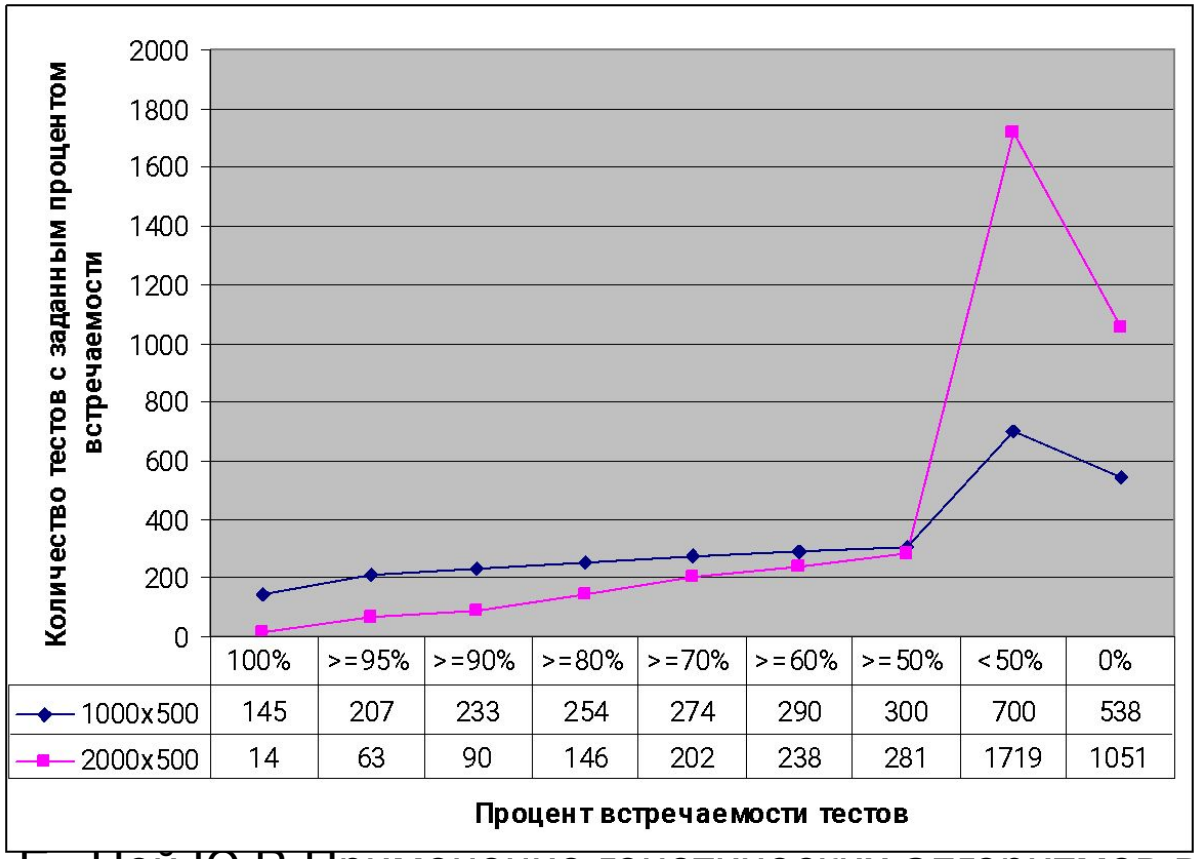
# 4. Результаты экспериментов

Несмотря на то, что увеличение размера популяции способствует повышению сходимости ГА, в соответствии с используемыми критериями, получены результаты, свидетельствующие 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.

В силу приведенного анализа результатов сокращение количества особей в популяции в $a_1$ раз и количества запусков ГА в $a_2$ раз, позволяет в ряде случаев уменьшить вычислительные затраты и время поиска решения пропорционально произведению $a_1 a_2$.

# 5. Модифицированный алгоритм

При увеличении количества строк в матрице тестов сходимость ГА существенно уменьшается.




| | 100% | >=95% | >=90% | >=80% | >=70% | >=60% | >=50% | <50% | 0% |
|---|---|---|---|---|---|---|---|---|---|
| 1000x500 | 145 | 207 | 233 | 254 | 274 | 290 | 300 | 700 | 538 |
| 2000x500 | 14 | 63 | 90 | 146 | 202 | 238 | 281 | 1719 | 1051 |

Янковская А.Е., Цой Ю.Р. Применение генетических алгоритмов в интеллектуальных распознающих системах.

# 5. Модифицированный алгоритм

Для повышения сходимости ГА предлагается следующая модификация с адаптацией к условиям эволюционного поиска.

Пусть $\nu(t)$ – количество ББДТ в матрице тестов в поколении $t$, $\nu(0) = n$, и $\nu'(t)$– количество ББДТ, не входящих в закодированные в популяции решения за последние  поколений и соответствующие неиспользуемым строкам из исходной матрицы тестов **T**.

Представим пошагово модифицированный ГА:

**Шаг 1.** Инициализация.

**Шаг 2.** Осуществить $\Delta t$ поколений эволюционного поиска.

**Шаг 3.** Если $\nu'(t)>0$ и $\nu(t) > n_0$, то удалить из матрицы **T** строку с минимальным суммарным весом и провести коррекцию $\nu(t+1) = \nu(t) - 1$.

**Шаг 4.** Если не выполняются условия останова ГА, то перейти на Шаг 2. Иначе переход на Шаг 5.

**Шаг 5.** Конец.

# 5. Модифицированный алгоритм

Однако полученные результаты экспериментов не выявили улучшений качества решений. В ряде случаев наблюдалось ухудшение результатов. В качестве возможного объяснения предполагается, что удаление неиспользуемых строк может либо случайно удалить «хорошую» строку (в случае, если строка удаляется на первых поколениях), либо не влияет на результат (если удаление происходит после наступления сходимости ГА).

Тем не менее, авторы надеются, что возможно улучшение алгоритма с удалением строк, т.к. поскольку увеличение количества строк приводит к ухудшению сходимости, то, вполне вероятно, что должен наблюдаться и «обратный эффект», когда уменьшение количества строк способствует повышению сходимости алгоритма.

# 6. Заключение

В докладе рассматривалось применение ГА для решения задачи формирования оптимального 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, поскольку необходимо проверить полученный результат на реальных данных.

# 6. Заключение

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

Программная реализация рассматриваемых алгоритмов создана с использованием инструментальной библиотеки классов Evolutionary Computation Workshop (http://qai.narod.ru/ecw/).

Планируется включение программного модуля, реализующего ГА в интеллектуальное инструментальное средство ИМСЛОГ.

# Благодарности

Исследование выполнено при поддержке грантов РФФИ (проект № 07-01-00452) и РГНФ (проект № 06-06-12603В)

# Список источников


1. Naidenova R.A., Plaksin M.V., Shagalov V.L. Inductive inferring all good classification test // Знание-Диалог-Решение. Сб. науч. тр. междунар.конф., том 1, Ялта, 1995. с.79-84.
2. Янковская 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. Часть I. -- Новосибирск, 1998. - С.195-199.
3. Yankovskaya A.E. Test Pattern Recognition with the Use of Genetic Algorithms // Pattern Recognition and Image Analysis, vol. 9, no. 1, 1999, p. 121-123.
4. Yankovskaya A.E. The Test Pattern Recognition with Genetic Algorithms Use // Proceedings of the Pattern Recognition and Image Understanding. 5th Open German-Russian Workshop. -- Germany, Herrshing, 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. ISSN 1561-535. Донецк, № 2, 2000,
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, vol. 9, no. 3(27), 2001, p. 336-349.
7. Yankovskaya A.E. Bleikher A.M. Optimization of tests synthesis on the base of descent algorithms with the use of genetic transformations // Radioelectronics & Informatics, no. 3(24), 2003, p. 51-55.
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 Stepanov V.V., 2005. P. 123-126.
9. Журавлев Ю.И., Гуревич И.Б. Распознавание образов и анализ изображений // Искусственный интеллект: В 3-х кн. Кн.2. Модели и методы: Справ. / Под ред. Д.А.Поспелова. М.: Радио и связь, 1990.
10. Янковская А.Е. Построение логических тестов с заданными свойствами и логико-комбинаторное распознавание на них // ИОИ-2002. Тез. докл. межд. науч. конф. -- Симферополь, 2002. -- С. 100-102.
11. Янковская А.Е., Цой Ю.Р. Исследование эффективности генетического поиска оптимального подмножества безызбыточных тестов для принятия решений // Искусственный интеллект. Научно-теоретический журнал, 2006, с. 257-260.

# Спасибо за внимание!

## Применение генетических алгоритмов в интеллектуальных распознающих системах

Янковская А.Е., Цой Ю.Р.
Томский государственный архитектурно-строительный университет
Томский политехнический университет
yank@tsuab.ruyank@tsuab.ru, qai@mail.ru