УДК 32.81

ГЕНЕТИЧЕСКИЙ АЛГОРИТМ И ЕГО МОДИФИКАЦИЯ ДЛЯ ФОРМИРОВАНИЯ ОПТИМАЛЬНОГО ПОДМНОЖЕСТВА ТЕСТОВ*

Ю.Р. Цой [1], А.Е. Янковская [2]

Рассматривается проблема выбора оптимального подмножества безызбыточных безусловных диагностических тестов (ББДТ) с использованием эволюционного подхода. Приводится генетический алгоритм и результаты исследования его эффективности для матриц псевдослучайных ББДТ. Показывается высокая сходимость алгоритма. Излагается модификация алгоритма, предназначенная для повышения качества результатов для матриц ББДТ большой размерности.

Введение

Формирование и выбор «хороших» [Naidenova, Plaksin, Shagalov, 1995] 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 [Янковская, 1998, Yankovskaya, 1999]. Первые алгоритмы построения ББДТ, описанные в [Янковская, 1998, Yankovskaya, 1999], программно реализованы и развиты в плане оптимизации построения в последующих работах Янковской А.Е. и Янковской А.Е. с Блейхер А.М. [Yankovskaya, Bleikher, 2003].

Однако, выбор «хороших» ББДТ не всегда приводит к оптимальному решению, поскольку общее количество признаков в выбранном множестве тестов может быть слишком большим, также как временные и стоимостные затраты или ущерб (риск) [Yankovskaya, Tsoy, 2005],

* Работа выполнена при финансовой поддержке РФФИ (проект № 07-01-00452) и РГНФ (проект № 06-06-12063В).

[1] *Томский политехнический университет, 634050, Томск, пр. Ленина, 30* qai@mail.ru

[2] *Томский государственный архитектурно-строительный университет, 634003, Томск, пл. Соляная, 2,* ayyankov@gmail.com

наносимый в результате выявления значений признаков исследуемого объекта, например, в медицине. Впервые критерии оптимальности и актуальная задача поиска оптимального подмножества безызбыточных безусловных диагностических тестов (ББДТ) поставлена в статье [Янковская, 2002]. В статье [Yankovskaya, Mozheiko, 2004] сформулированы критерии оптимальности, а в статье [Колесникова и др., 2005] предложено три алгоритма, обеспечивающие выполнение этих критериев: логико-комбинаторный, алгоритм на основе метода анализа иерархии и генетический алгоритм (ГА). В докладе представлен ГА [Янковская, Цой, 2008] и его модификация и результаты исследований.

1. Определения и обозначения

Воспользуемся определениями и обозначениями, необходимыми для постановки задачи и при дальнейшем изложении [Янковская, 2002, Журавлев, Гуревич, 1990].

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.

Пусть $\mathbf{T} = \{t_{ij} : i = 1,...,n,\ j = 1,...,m\}$ – матрица ББДТ, n – количество ББДТ, m – количество характеристических признаков, строкой $\mathbf{T}_i$ представлен i-й ББДТ. 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 [Yankovskaya, Tsoy, 2005]. Далее будем использовать термины «вес», «стоимость» и «ущерб» признака вместо соответственно «весовой коэффициент», «коэффициент стоимости» и «коэффициент ущерба».

Будем рассматривать случай бинарной матрицы **T** и 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) в $\mathbf{N}^0$ должно содержаться максимальное число псевдообязательных признаков; 2) $\mathbf{N}^0$ должно содержать минимальное общее число признаков; 3) $\mathbf{N}^0$ должно иметь максимальный суммарный вес; 4) $\mathbf{N}^0$ должно иметь наименьшую суммарную стоимость; 5) $\mathbf{N}^0$ должно иметь наименьший суммарный ущерб.

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 бинарной 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 $\mathbf{T}(h)$ в соответствии с выражением [Yankovskaya, Tsoy, 2005]:

$$f_k = \sum_{j=1}^{5} v_j e_h^{(j)} + 100\big(U(h) - n_0\big)^2, \qquad f \to \min,$$

где v_j – весовой коэффициент j-го критерия, соответствующий его значимости; $U(\boldsymbol{\psi})$ – количество единичных разрядов в бинарной строке $\boldsymbol{\psi}$; $e_h^{(j)}$ – функция штрафа за невыполнение j-го критерия:

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

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

где $S_W(\boldsymbol{\Psi}), S_{W'}(\boldsymbol{\Psi})$ и $S_{W''}(\boldsymbol{\Psi})$ – соответственно суммарный вес, стоимость и ущерб по всем тестам множества, соответствующего матрице $\boldsymbol{\Psi}$; $U_c(\boldsymbol{\Psi}) = U\left(\bigwedge_i \boldsymbol{\psi}_i\right)$ и $U_d(\boldsymbol{\Psi}) = U\left(\bigvee_i \boldsymbol{\psi}_i\right)$ – соответственно количество единичных разрядов в конъюнкции и дизъюнкции по всем строкам бинарной матрицы $\boldsymbol{\Psi}$.

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 применения ГА для решения поставленной задачи проведено с использованием псевдослучайных матриц тестов размерностями 1000x50, 1000x100, 1000x200, 1000x300, 1000x400, 1000x500, 2000x500. 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 равна 300.

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: $v_1 = 40$, $v_2 = 30$, $v_3 = 15$, $v_4 = 10$, $v_5 = 5$. Рассматривается ГА с турнирной селекцией с размером турнира равным 6, двухточечным оператором кроссинговера, битовой мутацией и 1 элитной особью. По итогам 100 независимых запусков для каждой из рассматриваемых матриц оцениваются результаты как по полученному лучшему значению функции приспособленности, так и по параметрам, сформулированным в [Янковская, Цой, 2006] и характеризующим стабильность решений, полученных в различных запусках:

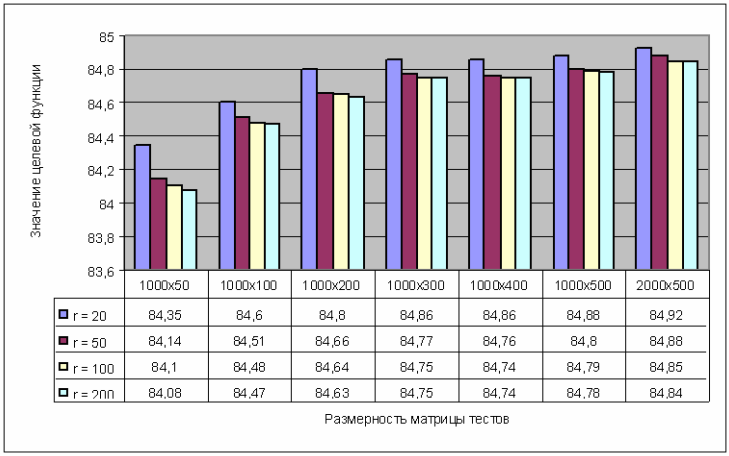
1. Критерий стабильности, учитывающий частоту p_i встречаемости i-го теста во всех решениях, полученных по результатам 100 запусков ГА. Чем больше количество тестов, для которых значение p_i равно или близко к 1, тем выше сходимость алгоритма.

2. Суммарное количество Ω ББДТ, не вошедших в полученные решения. Чем больше значение Ω, тем выше сходимость алгоритма.

Полученные лучшие значения целевой функции, усредненные по 100 запускам, для различных матриц ББДТ в зависимости от размера популяции показаны на рис. 1. При увеличении размера r популяции

отметим улучшение результатов, однако это улучшение весьма незначительно, в большинстве случаев, порядка 10^{-2}.

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 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.



	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

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

Также было проведено исследование зависимости состава подмножества тестов, сформированного по результатам нескольких запусков ГА, от количества запусков. 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. Таким образом, несмотря на различное количество запусков и размер популяции, 245 и 244 теста присутствуют в большинстве найденных решений.

Отметим, что увеличение размера популяции способствует повышению сходимости ГА по критериям из работы [Янковская, Цой, 2006], однако получены результаты, свидетельствующие о том, что для

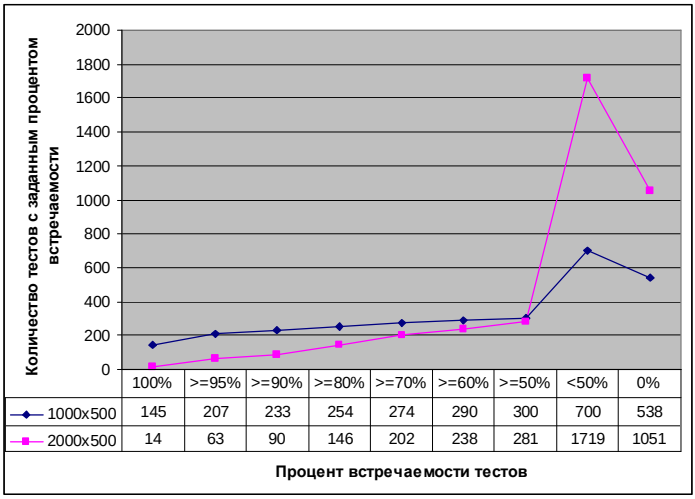
матриц тестов, имеющих не больше 1000 строк, анализ решений, сформированных при использовании сравнительно небольшого размера популяции и малого количества запусков, позволяет построить подмножество тестов, близкое к оптимальному.

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.

Таким образом, сокращение количества особей в популяции в a_1 раз и количества запусков ГА в a_2 раз, позволяет 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. Например, количество тестов, встречающихся во всех решениях по результатам 100 запусков для ГА с популяцией из 200 особей и матрицы тестов размером 1000x500, равно 145. При использовании матрицы размером 2000x500 количество тестов, встречающихся во всех решениях, уменьшается до 14. График, показывающий соотношение количества тестов в зависимости от частоты их встречаемости в полученных решениях показан на рис. 2.



	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

Рис. 2. Зависимость количества тестов от частоты их встречаемости в полученных решениях для популяции из 200 особей

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

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

Представим пошагово модифицированный ГА:

Шаг 1. Инициализация.

Шаг 2. Осуществить Δt поколений эволюционного поиска.

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

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

Шаг 5. Конец.

Таким образом, в модифицированной версии алгоритма осуществляется постепенное удаление неиспользуемых строк из исходной матрицы тестов, что позволяет сократить пространство поиска и должно обеспечить лучшую сходимость ГА по сравнению с первоначальной версией.

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 (рис. 2), то, вполне вероятно, что должен наблюдаться и «обратный эффект», когда уменьшение количества строк способствует повышению сходимости алгоритма.

Заключение

Полученные в результате исследований результаты показывают высокую эффективность эволюционного подхода к выбору оптимального подмножества ББДТ для псевдослучайных матриц тестов. Однако, приведенные выводы требуют проверки при решении практических задачах. Отметим, что использование предложенной модификации алгоритма, связанной с удалением неиспользуемых строк в исходной матрице тестов, в настоящее время не способствует повышению качества результатов. Сделаны предположения относительно природы полученных

результатов исследования.

Дальнейшая работа будет направлены на разработку более эффективных процедур эволюционного поиска оптимального подмножества ББДТ для решения задач принятия решений на основе тестового распознавания образов.

Список литературы

[Naidenova, Plaksin, Shagalov, 1995] Naidenova R.A., Plaksin M.V., Shagalov V.L. Inductive inferring all good classification test // Знание-Диалог-Решение. Сб. науч. тр. межд.конф., т. 1, Ялта, 1995. – С. 79-84.

[Янковская, 1998] Янковская А.Е. Тестовое распознавание образов с использованием генетических алгоритмов // Распознавание образов и анализ изображений: новые информационные технологии (РОАИ-4-98). Тр. IV Всерос. конф. с межд. уч. Ч. I. – Новосибирск, 1998. – С. 195-199.

[Yankovskaya, 1999] 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.

[Yankovskaya, Bleikher, 2003] 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.

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

[Янковская, 2002] 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.

[Yankovskaya, Mozheiko, 2004] Yankovskaya A.E., Mozheiko V.I. Optimization of a set of tests selection satisfying the criteria prescribed // 7th Int. Conf. PRIA-7-2004. Conf. Proc. Vol. I. – St. Petersburg: SPbETU 2004. – Pp.145-148.

[Колесникова и др., 2005] Колесникова С.И., Можейко В.И., Цой Ю.Р., Янковская А.Е. Алгоритмы выбора оптимального множества безызбыточных диагностических тестов в интеллектуальных системах поддержки принятия решений // Первая межд. конф. САИТ-2005: Тр. конф. В 2 т. Т.1. – М.: КомКнига, 2005. – С. 256-262.

[Журавлев, Гуревич, 1990] 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. – С. 149-191.

[Янковская, Цой, 2006] Янковская А.Е., Цой Ю.Р. Исследование эффективности генетического поиска оптимального подмножества безызбыточных тестов для принятия решений // Иск. интеллект. Науч.-теор. журн., 2006, с. 257-260.

[Янковская, Цой, 2008] Янковская А.Е., Цой Ю.Р. О применении генетических алгоритмов в интеллектуальных распознающих системах // Таврический вестник информатики и математики. – 2008. – № 2. – С. 262-270.