ОПТИМИЗАЦИЯ СИНТЕЗА БЕЗЫЗБЫТОЧНЫХ ДИАГНОСТИЧЕСКИХ ТЕСТОВ С ИСПОЛЬЗОВАНИЕМ ГЕНЕТИЧЕСКИХ АЛГОРИТМОВ И РЕАЛИЗАЦИЯ ЕЕ В ИНТЕЛЛЕКТУАЛЬНОЙ СИСТЕМЕ

А. Е. Янковская, А. М. Блейхер[1]
Работа поддержана РФФИ, проект №№ 98-01-00295, 98-01-03019.

Томский государственный архитектурно-строительный университет,
Лаборатория интеллектуальных систем
634003, Россия, Томск-3, Соляная пл.2
e-mail: yank@tisi.tomsk.su

[1]Томский политехнический университет,
Факультет автоматики и вычислительной техники, кафедра прикладной математики
636034, Россия, Томск-34, пр. Ленина, 30
e-mail: balm@am.tpu.ru, bleikher@chat.ru

ABSTRACT
Method optimization of syntheses a set of irredundant diagnostic tests with genetic algorithms use to solve tasks of large dimension is suggested. The idea of creating sectionalized by classification mechanisms of irredundant the partial implication matrix is lying in the base; revealing certain kinds of regularities, that are used, combined with genetic transformations, in creating a set of irredundant diagnostic tests. 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 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 tasks the intelligent system is used.

ВВЕДЕНИЕ
Задача оптимизации синтеза безызбыточных диагностических тестов возникает при решении задач распознавания образов в признаковом пространстве большой размерности. Проблеме распознавания образов посвящено довольно большое число исследований, многие из которых отражены в обзоре [1]. Генетические алгоритмы в распознавании образов стали интенсивно использоваться благодаря бурно развивающимся в последние годы работам в области мягких вычислений (термин мягкие вычисления введен Заде [2] в 1993 г.), к каковым и относятся генетические алгоритмы (ГА). Канонический генетический алгоритм описан в работе Holland I.H. [3]. Обзор работ по генетическим алгоритмам и их развитию приведен в [4,5], а применение генетических принципов в распознавании образов в [6].
Проводимые нами исследования по применению генетических алгоритмов для оптимизации безызбыточных диагностических тестов (БДТ) имеют небольшую историю. Первые результаты отражены в публикациях [7-9]. В публикации [7] описывается построение оптимальных смешанных диагностических тестов с использованием генетических преобразований на множестве псевдоальтернативных признаков (переменных), получаемых путем логико-комбинаторных процедур. При этом задача построения тестов сводилась к поиску всех минимальных (оптимальных) кратчайших столбцовых покрытий булевой матрицы импликаций, получаемой из матрицы описаний Q объектов в пространстве признаков и матрицы различений R, задающей различные механизмы разбиения объектов на классы эквивалентности. Оптимизация и генетические преобразования над матрицей импликаций основывались на применении операции кроссинговера на множестве признаков (генов), не входящих в число константных, обязательных и неинформативных и процеду-

ры отбора конкурентоспособных индивидов популяций (подматриц матрицы Q). В работе [8] приводится алгоритм построения безусловных безызбыточных диагностических тестов на основе безызбыточной матрицы импликаций с использованием генетических преобразований, а в работе [9] синтез безызбыточных диагностических тестов осуществляется без построения матрицы импликаций, а на базе шагово-циклического алгоритма, применяемого при кодировании внутренних состояний асинхронных автоматов [10].

Работы, описанные в публикациях [7-9], показали, что задача построения кратчайших покрытий успешно решалась при небольшом признаковом пространстве, когда число признаков не превышало 200-400. При увеличении признакового пространства возникали проблемы с машинными ресурсами (объем требуемой памяти и временные затраты), что стимулировало создание алгоритма, основанного на построении только части матрицы импликаций и применении шагово-циклического алгоритма [10], обеспечивающего сокращение переборных процессов при использовании генетических преобразований.

Синтез безызбыточных диагностических тестов с генетическими преобразованиями осуществляется путем частичного построения секционированной по механизмам классификации безызбыточной матрицы импликаций; выявления некоторых закономерностей в знаниях и нахождения весовых коэффициентов признаков; формирования популяций потомков, наследующих обязательные и часть информативных признаков родительской популяции, причем селекция из предыдущих популяций экземпляров с желаемыми свойствами производится по критерию минимизации числа кодируемых единичными значениями генов, входящих в хромосому (тест) и максимизации веса теста, вычисляемому на основе весовых коэффициентов признаков и являющемуся функцией стоимости. Приводится краткое описание интеллектуальной системы GenPro, предназначенной для оптимизированного синтеза БДТ.

ОСНОВНЫЕ ПОНЯТИЯ, СПОСОБ ПРЕДСТАВЛЕНИЯ ЗНАНИЙ И ФОРМИРОВАНИЯ ПОПУЛЯЦИЙ ПОТОМКОВ

Для понимания изложения приводятся основные понятия из публикаций [7-11].

Знания в интеллектуальной системе представлены в виде 2-х матриц: троичной матрицы описаний Q и целочисленной матрицы различений R.

Строкам матрицы Q сопоставляются описания объектов, столбцам - характеристические признаки. Элемент $q_{i,j}$ принимает значение “1”, если *j*-ый признак присущ *i*-му объекту, “0” – не присущ, “-” – значением признака может быть как 0, так и 1. Вес строки задается с помощью коэффициента p_i. Строкам матрицы R сопоставляются строки матрицы Q, столбцам - классификационные признаки, соответствующие различным механизмам классификации, разбивающим изучаемые объекты на классы эквивалентности. Элементы *j*-го столбца матрицы R задают номера классов, которым принадлежат объекты при *j*-ом механизме классификации. При этом считается, что объекты, которым соответствуют равные строки матрицы R, принадлежат одному образу, а множество соответствующих им строк матрицы Q задает описание данного образа.

Данная модель позволяет представлять не только данные, но и знания экспертов, поскольку одной строкой матрицы Q можно задавать в интервальной форме подмножество объектов, для которых характерны одни и те же решения, задаваемые строкой матрицы R. При обнаружении пересечения описаний образов, матрица Q должна быть доопределена.

Под закономерностями понимаются подмножества признаков с определенными легко интерпретируемыми свойствами, влияющими на различимость объектов из разных образов, устойчиво наблюдаемыми для объектов из обучающей выборки и проявляющимися на других объектах той же природы, а также весовые коэффициенты признаков, характеризующие их индивидуальный вклад в различимость объектов. К упомянутым подмножествам будем относить константные, устойчивые (константные внутри класса), неинформативные (не различающие ни одной пары объектов, но не константные или весовой ко-

эффициент признака меньше наперед заданного числа), альтернативные (в смысле включения в диагностический тест), зависимые (в смысле включения подмножеств различимых пар объектов), несущественные (не входящие ни в один безызбыточный тест), обязательные (входящие во все безызбыточные тесты) признаки.

Для представления условий различимости обучающих объектов используется секционированная по механизмам классификации двоичная матрица импликаций U, построенная по матрицам Q и R, столбцам которой сопоставляются характеристические признаки, а строкам – результаты сравнения всевозможных пар объектов, принадлежащих разным классам по каждому из механизмов классификации. Строка матрицы U представляет собой вектор-функцию различения. Компонента *m* вектор-функции различения принимает единичное значение (признак различающий), если *m*-ый признак в описании пары объектов принимает противоположное значение (0(1) - в первом, 1(0) - во втором описании объекта); и нулевое - в противном случае. Если только одна компонента вектор-функции различения равна единице, то признак, соответствующий этой компоненте, является обязательным и включается во все безызбыточные диагностические тесты.

Будем называть матрицу импликаций безызбыточной и обозначать далее через U', если в ней отсутствуют поглощающие строки. Алгоритм оптимизации ее построения приведен в статье [8].

Будем называть безызбыточную матрицу импликаций частичной и обозначать далее через $U'_ч$, если она задает только часть условий матрицы импликаций U'.

Ускоренное построение матрицы U' основано на упорядочении столбцов матрицы R по неубыванию числа пар объектов, подлежащих различению по механизмам классификации, и специальном порядке выбора сравниваемых пар объектов, упорядоченных (внутри классов по каждому механизму классификации) по неубыванию числа единичных значений в их описаниях. Поочередно сравниваются пары объектов из разных классов (с начала и конца описаний классов), что уменьшает (без потери закономерностей) число включаемых в матрицу U' строк, которые становятся поглощающими на более поздних этапах ее построения, и приводит к сокращению числа переборов при сравнении (на предмет поглощаемости) каждой новой строки с уже построенной частью матрицы U'.

В целях оптимизации использования памяти при имеющихся ресурсах строится U' с одновременным вычислением всех весовых коэффициентов признаков. При ограниченных ресурсах памяти строится $U'_ч$ с одновременным построением весовых коэффициентов w_m и последовательным построением вектор-функций различения и выделением обязательных и константных признаков.

Весовой коэффициент w_m признака, соответствующего *m*-му столбцу матрицы Q (*m*=1,...,M) вычисляется по формуле:

$$w_m = \frac{\sum_{r=1}^{K-1}\sum_{t=r+1}^{K}\sum_{i=1}^{N_r}\sum_{j=1}^{N_t}\delta_{ij}^m}{\sum_{i=1}^{K-1}\sum_{j=i+1}^{K}\sigma_i\sigma_j} \quad (1),$$

где K - число выделенных образов; N_f - число строк в описании *f*-го образа ($f \in \{r,t,i,j\}$); $\delta_{ij}^m = 0$, если $q_{i,m}=q_{j,m}=0$ или $q_{i,m}=q_{j,m}=1$ ($q_{i,m}$ - значение элемента матрицы Q, лежащего на пересечении *i*-ой строки и *j*-го столбца); $\delta_{ij}^m = p_i p_j 2^{d_i^- + d_j^-}$ (d_i^- - число значений "-" в *i*-ой строке матрицы Q, p_i - коэффициент повторения *i*-ой строки), если $q_{i,m}=0$ и $q_{j,m}=1$ или

$q_{i,m}$=1 и $q_{j,m}$=0; $\delta_{ij}^{m} = p_i p_j 2^{d_i^- + d_j^- - 1}$, если $q_{i,m}$="-" и (или) $q_{j,m}$="-"; σ_j - число объектов в *j*-ом образе (*j*=1,...,K), вычисляемое по формуле: $\sigma_j = \sum_{l=1}^{N_j} p_l 2^{d_l^-}$.

Если $w_m = 0$, то *m*-й признак является константным.

Будем считать *m*-й признак неинформативным, если он не является константным и если $w_m < c$, (*m*=1,...,M), где c – заранее заданная величина, определяемая из матриц Q и R.

Безызбыточные тесты строятся с использованием приводимых ниже генетических преобразований с включением обязательных признаков. Константные и неинформативные признаки в тест не включаются.

Функция стоимости диагностического теста зависит от длины теста и весовых коэффициентов признаков, входящих в тест. Минимальная длина диагностического теста равна:

$$L =]\log_2 K[\quad (2),$$

где K - число выделенных образов, а]b[- наименьшее сверху целое к b. Доказательство тривиально.

Вес (стоимость) W_i теста равен сумме всех весовых коэффициентов признаков *i*-го теста.

Экземпляр (индивид) популяции, представляющий собой подматрицу матрицы Q со столбцами, сопоставленными признакам, входящим в безызбыточный диагностический тест, считается конкурентоспособным при меньшем числе единичных генов (признаков), включенных в хромосому (тест), и при большей величине суммы весовых коэффициентов единичных генов, входящих в данную хромосому.

Величина конкурентоспособности зависит от конкретных матриц Q, R. Конкурентоспособный индивид считается годным для включения в популяцию.

Размер (мощность) популяции определяется числом входящих в нее индивидов.

Считается популяция тем перспективней, чем больше отношение суммы весовых коэффициентов единичных генов, входящих в объединение хромосом индивидов, к мощности (числу) единичных генов из этого объединения.

Формирование популяции потомков, наследующей признаки предыдущей родительской, основывается на применении процедуры скрещивания (кроссинговера) хромосом на множестве генов, сопоставленных необязательным признакам, входящим в безызбыточные диагностические тесты (хромосомы родительской популяции) и достраивания, при необходимости, каждого индивида некоторым количеством признаков (генов) из множества объединения признаков родительской популяции, не являющимися обязательными и уже включенными в строящийся индивид популяции. Необходимость достраивания может быть вызвана тем, что получаемый результат может не быть диагностическим тестом. Кроме того, тест проверяется на безызбыточность. Если он не удовлетворяет свойству безызбыточности, то исключается признак (признаки в общем случае), являющийся избыточным, и только после этого индивид включается в популяцию. Добавление и исключение признаков (изменение значения гена в хромосоме) реализуют операцию мутации, которая не применяется в алгоритме оптимизации синтеза БДТ. Размер популяции потомков зависит от числа генов, используемых для операции кроссинговера, и от матриц Q, R.

АЛГОРИТМ ОПТИМИЗАЦИИ СИНТЕЗА БЕЗЫЗБЫТОЧНЫХ ДИАГНОСТИЧЕСКИХ ТЕСТОВ

Алгоритм оптимизации синтеза безызбыточных диагностических тестов состоит из следующих этапов:

1. Выполнение процедуры построения безызбыточной матрицы импликаций U' при обеспеченности ресурсными возможностями и переход к алгоритму, приведенному в

[8]. В противном случае построение только части U', представляющей собой $U'_ч$, вычисление w_m и определение обязательных признаков.

1.1. Вычисление по матрицам Q, R вектор-функций различения. Построение U' с одновременным вычислением w_m, выделением обязательных признаков и удалением поглощающих строк. Если U' построена полностью, то применение алгоритма [8].

1.2. Запоминание $U'_ч$ и фиксация указателя t – точки обработанных условий различения в матрице R.

1.3. Последовательное построение вектор-функций различения, начиная с точки t с одновременным вычислением w_m и определением обязательных признаков и добавлением их к ранее найденным.

2. Построение начальной популяции.

2.1. Построение каждого индивида (подматрица матрицы Q) с использованием шагово-циклического алгоритма путем включения в хромосому генов, соответствующих обязательным признакам и добавления в нее генов (признаков, генерируемых с помощью генератора случайных кодов с учетом весовых коэффициентов w_m), обеспечивающих построение безызбыточного диагностического теста. При этом для ускорения процедуры построения используется $U'_ч$.

2.2. Выделение конкурентоспособных индивидов, сопоставленным всем минимальным тестам и безызбыточным с наибольшими значениями W_i, включением их в начальную популяцию размером, определяемым экспериментом. Окончание процедуры построения популяции зависит от исчерпания временных ресурсов, отведенных на вычисление, либо повторения генерируемых индивидов.

3. Построение следующей перспективной популяции на основе родительской с использованием операции кроссинговера и проверкой обеспечения безызбыточности диагностических тестов, сопоставленным хромосомам, включаемым в перспективную популяцию. Окончание генерации популяций определяется построением всех безызбыточных тестов, исчерпанием задаваемого числа популяций или временных ресурсов.

КРАТКОЕ ОПИСАНИЕ ИНТЕЛЛЕКТУАЛЬНОЙ СИСТЕМЫ

Интеллектуальная система GenPro, создаваемая для оптимизации синтеза безызбыточных тестов с использованием генетических алгоритмов, реализуется на языке Borland C Builder для операционных систем семейства Windows95/NT.

Интеллектуальная система включает в себя следующие компоненты: инициализация, процесс обработки, представление результатов.

Инициализация состоит из процесса построения базы знаний и параметров, задаваемых для управления процессом генерации популяций, в частности, временных затрат и максимального числа допустимых популяций.

Для описания знаний в системе применяется структура PRows, задающая одной строкой строки матриц описаний Q и различений R. Структура состоит из следующих полей:

Name - название объекта (строки);

Item – строка матрицы Q, составленной как трехмерный массив, каждый элемент которого принадлежит множеству {0,1,2}, где 0 соответствует 0 в матрице Q, 1 - соответствует 1 в матрице Q, а 2 - соответствует “-“ в матрице Q;

Count - число столбцов (признаков) матрицы Q;

RItem - строка матрицы R аналогично построению строки матрицы Q;

RCount - число столбцов (различающих признаков) матрицы Q;

NImage - номер выделенного образа.

Каждый новый объект описывается в виде структуры PRows, а затем добавляется в переменную FQobj, являющуюся указателем на класс TList из библиотеки VCL. FQobj позволяет хранить всю информацию о матрицах Q и R.
Процесс обработки реализован в системе в виде класса TGenClass, который описывает все процедуры, выполняющие следующие операции: вычисление веса признаков (Computewm) генерация популяций (Generation), отбор конкурентоспособных индивидов (GetBestInd) и проверка на безызбыточность (IsIrreduntant).
Следует отметить, что процедура вычисления веса признаков проходит за одну итерацию и включает в себя процедуры построения безызбыточной матрицы импликаций, либо частичной матрицы импликаций (GetIr), нахождения обязательных и константных признаков.
Модуль представления результатов содержит процедуры сохранения и вывода на печать полученных результатов.

ЗАКЛЮЧЕНИЕ

Предложенный подход, реализованный в интеллектуальной системе, позволяет оптимизировать синтез БДТ при решении задач распознавания образов большой размерности.
Оптимизация достигается за счет частичного построения безызбыточной матрицы импликаций, используемой для генерации конкурентоспособных индивидов и популяций, построением за одну итерацию весовых коэффициентов генов и определением обязательных генов, включаемых во все хромосомы, а также применением генетических преобразований.
Есть все основания полагать, что реализованный в интеллектуальной системе метод оптимизации синтеза безызбыточных диагностических тестов с использованием генетических алгоритмов получит широкое распространение, поскольку он ориентирован на решение задач большой размерности, его параметры зависят от размерности задачи, ведется целенаправленный поиск перспективных индивидов и популяций.
Наличие интеллектуальной системы, реализованной на IBM PC/AT в среде Windows и основанной на тестовом распознавании образов, позволяет с относительно небольшими затратами дополнять ее процедурами, связанными с использованием других алгоритмов, тестировать их на уже созданных базах знаний из различных проблемных областей (медицина [11], медицина чрезвычайных ситуаций, геология, экология, генетика и др.).

ЛИТЕРАТУРА

1. Журавлев Ю.И., Гуревич И.Б. Распознавание образов и анализ изображений// Искусственный интеллект: Кн. 2. Модели и методы/ Под ред. Д.А.Поспелова. М.: Радио и связь, 1990. С. 149-190.
2. Zadeh L.A. Fuzzy logic, neral network and soft computing// Communication of the A.C.M. 1994. Vol. 37, № 3. P. 77-84.
3. Holland J.H. Adaptation in Natural and Artificial Systems// Ann Arbor: University Michigan Press, 1975.
4. Скурихин А.М. Генетические алгоритмы// Новости искусственного интеллекта. 1995. № 4. С. 6-46.
5. Курейчик В.М. Генетические алгоритмы. Состояние. Проблемы. Перспективы// Теория и системы управления М.: Наука, №1, 1999, С. 144-160.
6. Shmerko V., Yanushkevich S., Zaitseva E. Genetic Principles in Pattern Recognition and Image Processing.// Proc. of the Third International Conf. "Pattern Recognition and Information Analysis. Vol.3. Minsk - SZCZECIN, 1995. P. 17-24.
7. Yankovskaya A.Ye. Design of Optimal Mixed Diagnostic Test with Reference to the Problems of Evolutionary Computation// Proc. of the First International Conf. on Evolutionary Computation and Its Applications. “EvCA’96”. Moscow, 1996. P. 292-297.

8. Yankovskaya A.E.. The Test Pattern Recognition with Genetic Algorithm Use// 5th Open German-Russian Workshop on Pattern Recognition and Image Understanding./B. Radig, H. Niemann, Y. Zhuravlev, I. Gourevitch, I. Laptev (Eds.). – Germany, Herrshing, 1999. pp.47-54.
9. Yankovskaya A.E. The Test Pattern Recognition with Genetic Algorithm Use// Pattern Recognition and Image Analysis. – 1999.-Vol.9.No.1.-pp.121-123.
10.Янковская А.Е. Алгоритмы кодирования внутренних состояний асинхронного автомата//Сб. Цифровые модели и интегрированные структуры. Таганрог, 1970, с.390-399.
11.Янковская А.Е. Тестовые распознающие медицинские экспертные системы с элементами когнитивной графики// Компьютерная хроника. 1994. №№ 8/9. С. 61-83.