УДК 519.71

КОМПЬЮТЕРНАЯ СИСТЕМА ДЛЯ РЕШЕНИЯ ЗАДАЧ КЛАССИФИКАЦИИ НА ОСНОВЕ МОДИФИЦИРОВАННЫХ ИММУННЫХ АЛГОРИТМОВ

Литвиненко В.И., Дидык А.А., Захарченко Ю.А.

Введение. Искусственные иммунные системы (ИИС) являются новым направлением в исследованиях вычислительного интеллекта (ВИ). В настоящее время существует определенное число моделей ИИС, которые используются для решения задач распознавания образов, обнаружения отказов, компьютерной безопасности и многих других приложений [1]. Среди различных механизмов биологической иммунной системы при разработке ИИС наиболее часто используются модель иммунной сети и клональный отбор [2].

Целью работы является разработка компьютерной системы для решения задач распознавания и классификации с учителем, предназначенной для исследования свойств алгоритмов клонального отбора и искусственной иммунной сети.

Искусственные иммунные системы для решения задач распознавания. В настоящее время существует ряд компьютерных систем предназначенных для решения задач классификации на основе иммунных алгоритмов. К наиболее известным можно отнести WEKA (Waikato Environment for Knowledge Analysis) [5]. Система, основанная на использовании нейронной сети и искусственной иммунной сети для решения задач классификации (*http://wekaclassalgos.sourceforge.net/*), также включает реализованный на данной платформе Джейсоном Браунли [6] алгоритм клонального отбора. Система реализована на языке Java.

В предлагаемой авторами системе реализована способность иммунной системы распознавать и обучаться для решения задач классификации и распознавания.

Постановка задачи классификации. Для формальной постановки задачи классификации вектор признаков $x \in X \subseteq R^n$ является описанием объектов, которое является модулем преобразования. Классом называется некоторое подмножество $K_y = \left\{ x \in X \middle| y^*(x) = y \right\}$ множества X. Обучающей выборкой называется набор $T = (x_1, y_1), ..., (x_l, y_l)$, для которых $y^*(x_i) = y_i, i = \overline{1,l}$, то есть это известная информация об отображении $X \xrightarrow{y^*} Y$. Задача классификации заключается в построении функции классификации $F(x)$, приближающей отображение y^*, основываясь на обучающей выборке $(x_1, y_1), ..., (x_l, y_l)$.

В ИИС антитела (детекторы) и антигены (данные) имеют формальное представление в виде векторов координат (атрибутов): $Ab = (Ab_1, Ab_2, ..., Ab_l)$ для антитела и $Ag = (Ag_1, Ag_2, ..., Ag_l)$ для антигена. Без потери общности допустим, что данные вектора имеют одинаковый размер. В этом случае под аффинностью связей антител друг с другом или антител с антигенами понимается расстояние между соответствующими векторами атрибутов, которое выражается в виде скалярного неотрицательного значения: $P^l \times P^l \to \Re^+$, определяющего степень соответствия между молекулами (векторами атрибутов). Причем, при использовании оговоренной выше степени подобия форм, получается, что чем меньше расстояние между индивидуумами, тем выше их аффинность друг к другу. Значение расстояния может быть вычислено при помощи любой из приведенных ниже метрик:

- евклидово расстояние D_E (используется при вещественном или целочисленном кодировании атрибутов)

 ISBN 7-776-8361-7

$$D_E = \sqrt{\sum_{i=1}^{l} \left(Ab_i - Ag_i\right)^2}\ ; \qquad (1)$$

- манхэттенское расстояние D_M (также используется при вещественном или целочисленном кодировании)

$$D_M = \sum_{i=1}^{l} \left|Ab_i - Ag_i\right|. \qquad (2)$$

Более детально теория клонального отбора и искусственной иммунной сети, их формальное представление и алгоритмы описаны в работах [2,3,4].

Модифицированный алгоритм клонального отбора. Пошаговая реализация алгоритма клонального отбора представлена на рис.1. Основным отличием данной реализации алгоритма от классической для решения задач классификации есть оператор мутации. В связи с большим объемом данных при решении задач распознавания мы столкнулись с проблемой быстроты работы алгоритма.

1.Создание первоначальной популяции Ab
2.Вычисление аффинности каждого Ab:
3. Клонирование пропорционально аффинности
(чем выше аффинность, тем больше клонирование)
4. Гипермутация:
  4.1. Вычисление весов каждого гена антитела
  4.2. Сопоставление веса гена с заданным
    весовым порогом
5.Модифицированная гипермутация получение матрицы M_C
6. Расчет аффинности между модифицированными клетками
памяти M_C и каждым ag_i из Ag
7. Ранжирование антител из Ab по убыванию их аффинности.
8. Отбор n наилучших антител Ab' в результирующее множество
M_R
9. Генерация n_C новых антител $Ab_{Ген}$
10. Получение новой популяции $Ab_M = M_R \cup Ab_{Ген}$
11. Проверка критерия останова

Рис. 1 Алгоритм клонального отбора для решения задач классификации

Для ускорения процесса мутации было предложено для каждого гена антитела устанавливать весовой коэффициент. Если весовой коэффициент данного гена ниже, чем выбрано в установках построения сети, то ген необходимо подвергать мутации, в противном случае, оставлять как есть. Для установления весового коэффициента используются следующие формулы:

$$VS[ab_i] = \frac{\sum_{k=1}^{L} \left(Ab_k - Ag_k\right)^2}{\sqrt{\sum_{j=1}^{L} \left(Ab_j - Ag_j\right)^2}}, \qquad k \neq i \qquad (3)$$

$$VS[ab_i] = \frac{\sum_{k=1}^{L} |Ab_k - Ag_k|}{\sum_{j=1}^{L} |Ab_j - Ag_j|}, \qquad k \neq i \tag{4}$$

Формула (3) используется в случае, если расстояние между антителом и антигеном вычисляется как Евклидово расстояние, формула (4) – если расстояние вычисляется как Манхеттеновское расстояние. Весовой коэффициент назначается гену на каждом шаге мутации: ab_i – ген антитела Ab_j, для которого определяется вес; ab_j – ген антитела Ab_j; ag_j – ген антигена Ag_j; ab_k – ген антитела Ab_j, который не соответствует гену ab_i (т.е. $ab_i \in \lfloor Ab_j \rfloor$; $[Ab_k] \subset \lfloor Ab_j \rfloor$; $ab_i \notin [Ab_k]$). С целью оптимизации процесса обучения иммунной сети и формирования клеток памяти, а также для ускорения работы алгоритма клонального отбора, для формирования пула клонов при выполнении операции клонирования использовался фактор умножения. В нашей реализации *n* антител с самой высокой афинностью перед процессом клонирования сортировались в порядке возрастания. Таким образом, общее количество клонов, сгенерированных для всех этих *n* выбранных антител, устанавливалась в соответствии с формулой (5):

$$N_c = \sum_{i=1}^{n} round(\frac{\beta * N}{i}), \tag{5}$$

где N_c – общая сумма клонов, произведенных для каждого из антигенов; β - фактор умножения; N – общая сумма антител; *round* – оператор, который округляет аргумент к самому близкому целому числу. Каждый слагаемый этой суммы соответствует размеру клона каждого отобранного антитела, например, для N=100 и β=1, антитело с самой высокой афинностью (i=1) производит 100 клонов, в то время как второе по афинности антитело производит 50 клонов и т.д.

Модифицированный алгоритм иммунной сети. Иммунная сеть математически может быть представлена в виде графа, причем необязательно полносвязного, который состоит из множества узлов - клеток сети (антител) и множества взвешенных ребер, означающих связи между клетками. Значение веса ребра соответствует аффинности связи клеток друг с другом. В иммунных сетях различают два вида аффинности:

- аффинность связи «антиген-антитело» (*Ag-Ab*) – степень различия;
- аффинность связи «антитело-антитело» (*Ab-Ab*) – степень подобия.

Структура компьютерной системы. Блок интерфейсов. Включает интерфейс оператора (системного программиста). Данный блок выполняет следующие основные функции:

- ввод данных;
- вывод информации о результатах тестирования;
- управление процессом тестирования;
- настройку параметров подсистемы обучения искусственной иммунной сети.

Блок формирования искусственной иммунной сети для решения задачи классификации. Данный блок получает на вход записи, содержащие значения индивидуумов для конкретной задачи классификации. В соответствии с заданными оператором настройками и выбранным алгоритмом обучения, осуществляет формирование пула памяти искусственной иммунной сети согласно входному пулу индивидуумов.

Блок обработки результатов обучения ИИС. Осуществляет формирование статистической отчетности по результатам обучения искусственной иммунной сети на всех его этапах и созданию пула клеток памяти. К такой отчетности относится статистика созревания аффинности на каждой итерации обучения сети, количество клеток памяти сети и т.д.

 ISBN 7-776-8361-7

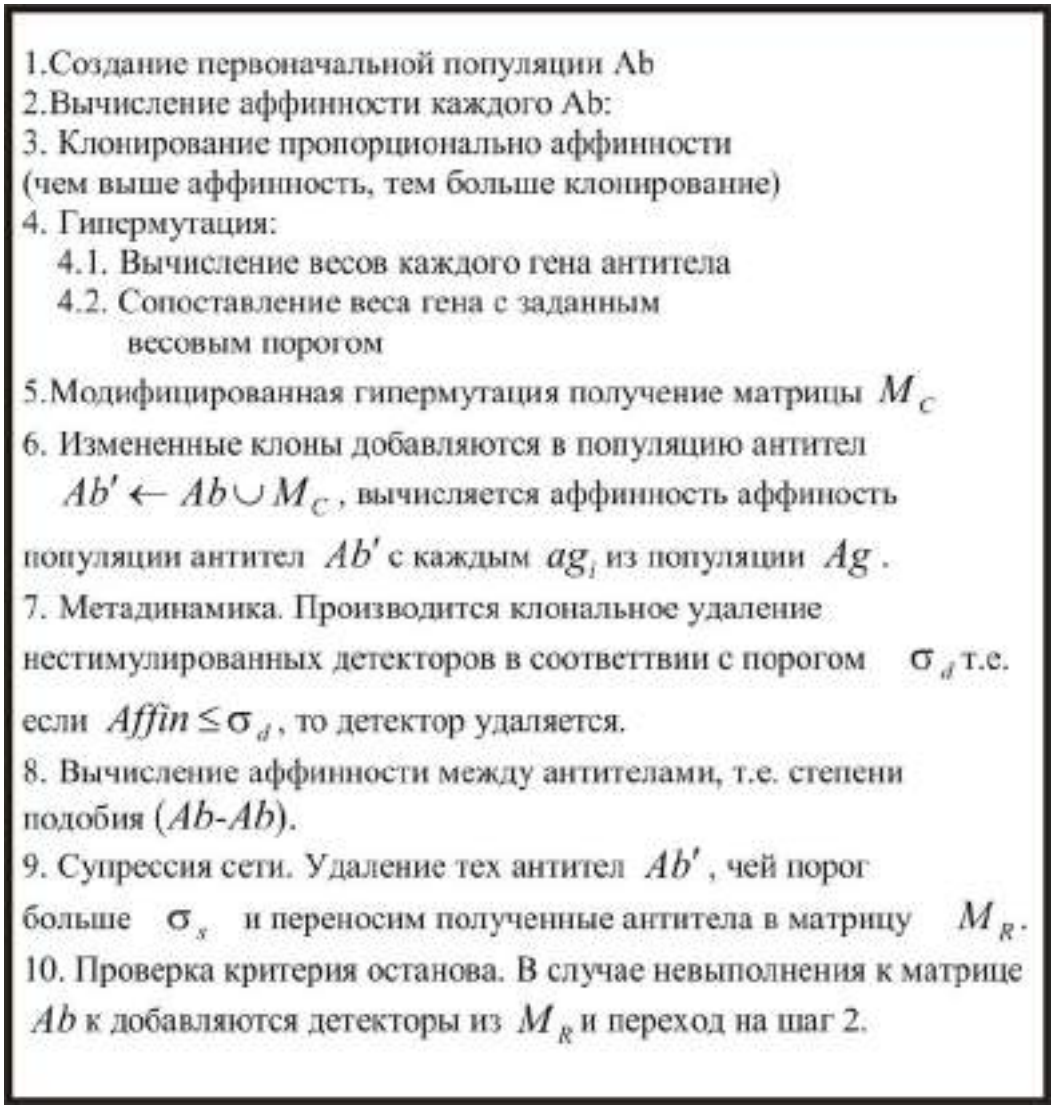

Рис. 2 Алгоритм искусственной иммунной сети для решения задач классификации.

База данных ИИС. База данных хранит следующие элементы:

- клетки памяти искусственной иммунной сети для соответствующей задачи классификации;
- статистические данные, полученные в процессе обучения искусственной иммунной сети.

Блок диспетчера тестирования ИИС. Блок имеет следующие функциональные характеристики:

- загрузка тестирующих записей из внешней базы данных;
- прогонка работы ИИС на загруженных данных;
- формирование отчетности по результатам решения задачи классификации.

Модульная структура системы. На рис. 4 показана модульная структура ИС для решения задачи классификации.

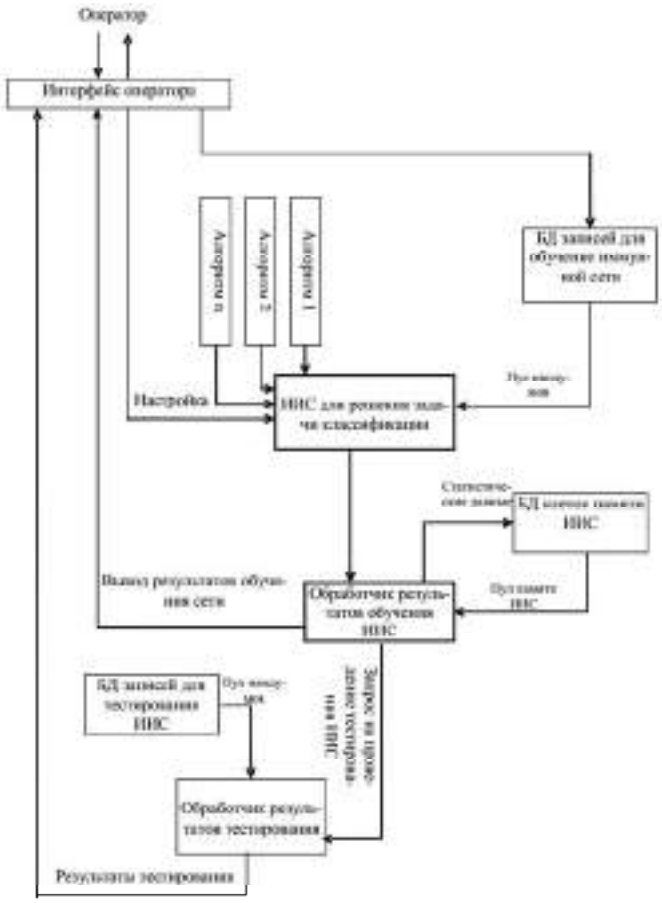

Рис. 3 Концептуальная структура компьютерной системы

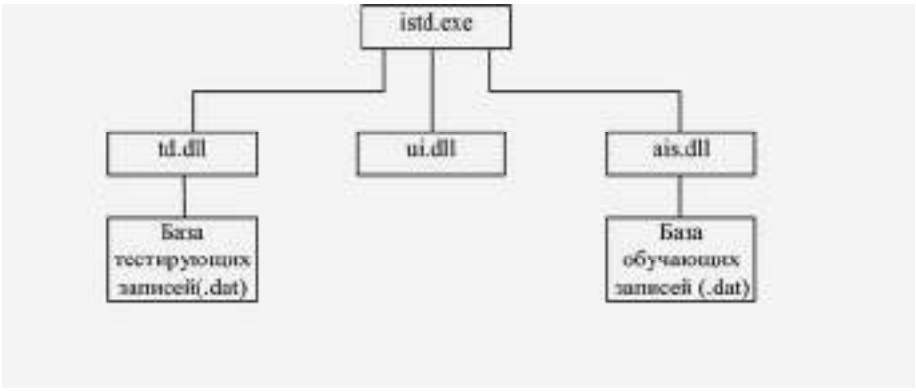

Рис. 4 Модульная структура системы

Модуль ais.dll содержит библиотеку классов ИИС. Каждый контролируемый параметр системы использует отдельный экземпляр ИИС. Модуль td.dll выполняет задачу проведения проверки работы ИИС на решение задачи классификации. В его функции входит получение и обработка записей для классификации из внешней базы данных и формирование результатов тестирования. Данные накапливаются в базе и выбираются оттуда по мере необходимости посредством запросов из модуля ais.dll. Модуль ui.dll является графическим интерфейсом ИС. Для хранения баз данных используются электронные таблицы в формате CSV.

Функционирование программы и параметры настройки. Используются следующие параметры для обучения :

 ISBN 7-776-8361-7

The size of population – размер популяции антител, формируемый на каждом шаге итерации для каждого индивидуума из обучающей выборки;

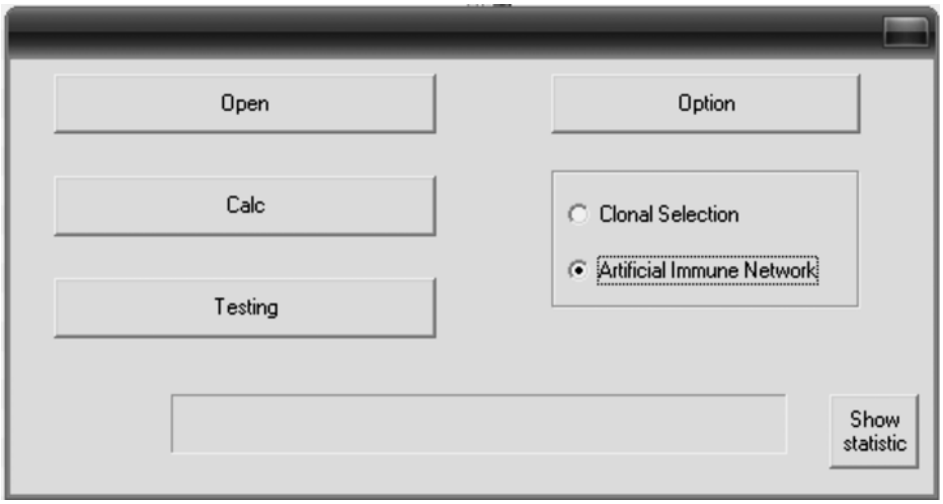

Рис.5 Общий вид интерфейса системы

Number of the best individums – число антител с наилучшей афинностью из популяции антител, которое не будет заменено на следующем шаге итерации случайными числами. Расачитывается как процентное соотношение от размера популяции антител.

Number of generations – число генераций для каждого антитела из обучающей выборки.

Mutation factor – фактор мутации антитела, который влияет на скорость приближения генерируемого антитела к обучающему. Допускается значение фактора в пределе 1-30. Чем больше значение фактора мутации, тем больше значение, на которое изменяется антитело на шаге мутации:

$$\alpha(D^*) = \exp(-\rho D^*) \qquad \textbf{(5)}$$

где ρ – параметр контролирующий сглаживание инверсии экспоненциала, D^* - норамлизированная афинность, расчитанная как $D^*=D/D_{max}$.

Input e – нормированный порог аффинности.

Mutation percent – параметр, задающий процент гипермутации антитела. Задается в процентном соотношении. Указывает число генов антитела, которые подлежат мутации. Количество изменяемых генов расчитывается следующим образом:

$$MG = round(MP/100) \qquad \textbf{(6)}$$

где MG – количество изменяемых генов, MP – заданный процент от общего количества генов антитела, *round* – оператор, который округляет аргумент к самому близкому целому числу.

Beta koef – фактор умножения для операции клонирования. Общее количество клонов, сгенерированных для всех n выбранных антител, устанавливалась в соответствии с формулой (5).

Mutation updating – включатель процедуры модифицированной мутации антител. Для ускорения процесса мутации было предложено для каждого гена антитела устанавливать весовой коэффициент формулы (3) и (4).

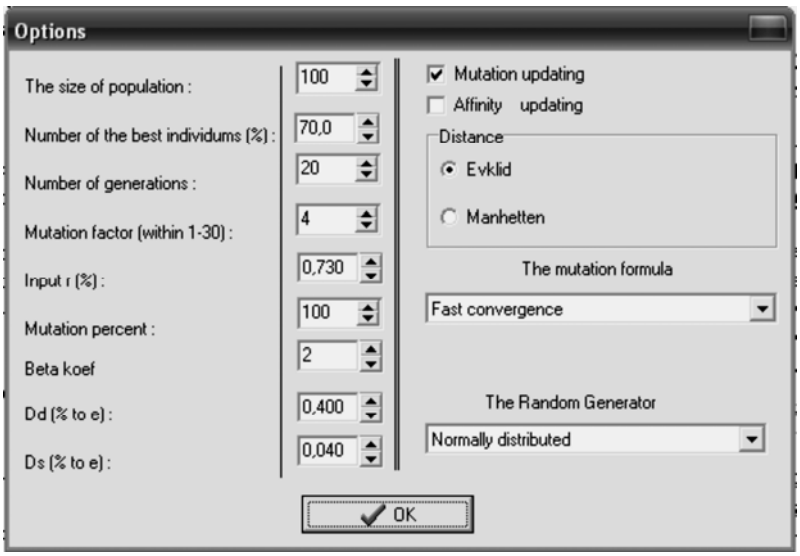

Рис.6 Интерфейс настройки основных параметров клонального алгоритма и искусственной иммунной сети

Affinity updating – включатель процедуры модифицированного расчета афинности антител. Афинность расчитывается по следующей формуле:

$$Aff_{ij} = \begin{cases} Dist_{ij} + \dfrac{n_j}{N_j}, & n_{\bar{j}} = 0 \\ 2 & n_{\bar{j}} > 0 \\ 0 & \end{cases} \qquad \textbf{(10)}$$

где Aff_{ij} - афинность между *i*-м антителом и *j*-им антигеном; $Dist_{ij}$ - расстояние между *i*-м антителом и *j*-им антигеном (нормированное от 0 до 1: 0 – максимально возможное расстояние, 1 – полное совпадение координат); n_j - количество антигенов, попавших в область охвата *i*-ого антитела, у которых ключевой класс совпадает с *j-ым* антигеном; N_j - общее количество антигенов, в которых ключевой класс совпадает с *j-ым* антигеном; $n_{\bar{j}}$ - количество антигенов, попавших в область охвата *i*-ого антитела, у которых ключевой класс не совпадает с *j-ым* антигеном.

Distance – вычисление расстояний между антителами:

Evklid – Эвклидово расстояние (1); Manhetten – Манхеттеновское расстояние (2).

The mutation formula – формула, по которой производится мутация антитела. Имеет следующие варианты выбора:

– *Fast convergence* : $c_j^* = c_j - \alpha(ab_i - c_j), \qquad j = 1,...,|C|;$

– *Affinity proportional mutation*: $c_j^* = c_j - \exp(-\alpha D^*), \qquad j = 1,...,|C|;$

– *Somatic mutation* : $c_j^* = c_j + \exp(-\alpha D) * N(0,\alpha), \qquad j = 1,...,|C|;$

где c_j – мутируемый антиген, α – коэффициент мутации, *N(0, α)* – функция случайного распределения в диапазоне (0, α); Td (σ_d) – порог смертности для антител в клетках памяти иммунной сети; Ts (σ_s) – порог суппрессии.

Выводы. Разработаны модифицированные алгоритмы и компьютерная система, предназначенные для решения задач классификации и исследования свойств алгоритмов для решения задач классификации на основе клонального отбора и искусственной иммунной сети. Система позволяет исследовать влияние операторов замещения, суппрессии, гипермутации и их модификаций на скорость сходимости и качество классификации. При-

 ISBN 7-776-8361-7

менение и исследование разработанной иммунной системы позволит создавать и разрабатывать алгоритмы классификации и распознавания образов для решения различных прикладных задач.

In article the computer system for the decision of problems of classification on the basis of the modified algorithms of an immune network and algorithm clonal selection is described. The architecture of system, program modules and parametres of adjustment of algorithms of training is described.

1. De Castro, L. N. & Timmis, J. I. Artificial Immune Systems: A New Computational Intelligence Approach, , London: Springer-Verlag 200o), September, 357 p.

2. Бідюк П.І. Литвиненко В.І. Фефелов. А.О. Формалізація методів побудови штучних імунних мереж// Наукові вісті НТУУ “КПІ”. 2007 р.–С.29-41

3. Литвиненко В.И. Искусственные иммунные системы как средство индуктивного построения оптимальных моделей сложных объектов // Проблемы управления и информатики. –. 2008.– № 3.– с. 43-61.

4. Литвиненко В. И. Иммунный классификатор для решения задач бинарной классификации (теоретические основы) //Системні технології. Регіональний міжвузівський збірник наукових праць. Випуск 1(42). –Дніпропетровськ, 2006. с.114-130.

5. http://sourceforge.net/projects/weak

6. Brownlee J. Clonal Selection Theory & CLONALG - The Clonal Selection Classification Algorithm (CSCA) // Victoria, Australia: Centre for Intelligent Systems and Complex Processes (CISCP), Faculty of Information and Communication Technologies (ICT), Swinburne University of Technology; 2005 Jan; Technical Report ID: 2-01.