УДК 004.89

ИССЛЕДОВАНИЕ ИММУННЫХ ОПЕРАТОРОВ В ЗАДАЧЕ КЛАСТЕРИЗАЦИИ ОБЪЕКТОВ

Н.М. Кораблев[1], А.А. Фомичев[2]

[1] ХНУРЭ, г. Харьков, Украина, korablev@kture.kharkov.ua
[2] ХНУРЭ, г. Харьков, Украина, alexandros_1985@mail.ru

В данной работе исследуются иммунные операторы, которые используются для решения задачи кластеризации объектов. На этапе обучения применяется приоритетное последовательное клонирование с конкурентно-целевым отбором клонов, и предложен метод разброса объектов в определяемых ограниченных областях. Спецификой работы иммунных операторов является использование единого критерия отбора объектов, применяемого как на этапе обучения, так и на этапе формирования кластеров.

ИСКУССТВЕННЫЕ ИММУННЫЕ СИСТЕМЫ, ПРИОРИТЕТНОЕ ПОСЛЕДОВАТЕЛЬНОЕ КЛОНИРОВАНИЕ, КОНКУРЕНТНО-ЦЕЛЕВОЙ ОТБОР КЛОНОВ, ОГРАНИЧЕННЫЕ ОБЛАСТИ РАЗБРОСА

Введение

В настоящее время в информационных технологиях используются подходы, позволяющие проводить интеллектуальную обработку и систематизацию данных. Одним из методов систематизации является кластеризация объектов, основной задачей которой является разбиение некоторого множества объектов по ряду признаков на кластеры [1-3]. Среди биологических принципов кластеризации можно выделить искусственные иммунные системы [4-7], рабочими элементами которых являются популяции антигенов и антител.

В настоящее время разработаны иммунные алгоритмы кластеризации, которые на некоторых этапах работы используют одинаковые подходы для решения задач клонирования, отбора и дополнительного разброса [6, 7]. В этих алгоритмах на этапах отбора антител и клонов производится большое количество избыточных вычислений. Это связано с тем, что клонируемые антитела конкурируют между собой за количество клонов. Такой способ организации клонирования не обеспечивает максимального покрытия антигенов, что увеличивает количество шагов в цикле их восстановления.

В предлагаемом алгоритме реализуются новые подходы, используемые в работе иммунных операторов клонирования и отбора клонов. Для отбора объектов используется единый критерий, основанный на плотности распределения исходной популяции антигенов, предложен новый способ организации клонирования, при котором каждое антитело создает максимально возможное количество клонов, в результате чего восстанавливается большее количество антигенов. Оператор отбора клонов реализуется на основе использования конкурентно-целевого отбора. Оператор дополнительного разброса реализуется путем разброса антител в определяемых ограниченных областях. Предложен также подход к определению центров кластеров и их границ. Использование этих методов сокращает избыточные вычисления и способствует ускорению работы иммунного алгоритма.

1. Постановка задачи

Дано множество объектов $M\{m_1;...;m_n\}$, $n=\overline{1,N}$, каждый из которых описывается набором признаков $S\{s_1;...;s_k\}$, $k=\overline{1,K}$ (в простейшем случае это координаты объекта). Эти объекты представляются в виде популяции антигенов $AG(ag_1;...;ag_n)$. В качестве меры близости между объектами m_i и m_j используется критерий аффинности Af_{ij} [4-7]:

$$Af_{ij}=(1+d_{ij})^{-1}, \tag{1}$$

$$d_{ij}=\sqrt{\sum_{m=1}^{k}(s_{im}-s_{jm})^2}, \tag{2}$$

где d_{ij} – евклидово расстояние между объектами; s_{im} – m-й признак у i-го объекта.

Необходимо разработать ИИС-алгоритм кластеризации исходного множества объектов $M\{m_1;...;m_n\}$, использующий в качестве основной меры близости объектов критерий (1), позволяющий производить их кластеризацию.

2. Алгоритм кластеризации

Работу иммунного алгоритма кластеризации можно разделить на два основных этапа: 1) обучение (восстановление популяции антигенов); 2) формирование кластеров.

В соответствии с этим процесс кластеризации может быть описан следующим образом:

$$\begin{gathered} fclust(ag_1;...;ag_n)=result(reconstr(ab_1;...;ab_n),\\ clust(ab_1';...;ab_n')),\\ (ab_1';...;ab_n')=reconstr(ab_1;...;ab_n), \end{gathered} \tag{3}$$

где $reconstr(ab_1;...;ab_n)$ – восстановление антигенов (обучение) путем применения операторов отбора, клонирования, мутации и старения к популяции

антител $AB(ab_1;...;ab_n)$, а $clust(ab_1';...;ab_n')$ – формирование кластеров. Тогда процесс восстановления антигенов может быть представлен следующим образом:

$$\begin{gathered} reconstr(ag_1;...;ag_n) = reconstr(sel(ab_1;...;ab_n), \\ clon(ab_1';...;ab_n'), ageing(ab_1'';...;ab_n''), \\ dispers(ab_1;...;ab_n)), \\ (ab_1';...;ab_n') = sel(ab_1;...;ab_n), \\ (ab_1'';...;ab_n'') = clon(ab_1';...;ab_n'), \end{gathered} \quad (4)$$

где $sel(ab_1;...;ab_n)$ – оператор отбора антител; $clon(ab_1';...;ab_n')$ – оператор клонирования, мутации и отбора клонов; $ageing(ab_1'';...;ab_n'')$ – оператор старения антител; $dispers(ab_1;...;ab_n)$ – оператор дополнительного разброса объектов.

В работе оператора отбора антител $sel(ab_1;...;ab_n)$ (первичного и после дополнительного разброса) используется критерий средней аффинности антигенов *CSel* [9]:

$$CSel = \frac{\sum_{i=1}^{n} AF_{iAG}}{n}, \quad (5)$$

где AF_{iAG} – средняя аффинность антигена ag_i со всеми антигенами выборки $AG(ag_1;...;ag_n)$, которая определяется следующим образом:

$$AF_{iAG} = \frac{\sum_{j=1}^{n} af_{ij}}{n}, \quad (6)$$

где af_{ij} – аффинность (1) антигена ag_i и антигена ag_j. Это упрощает процедуру отбора и повышает вероятность восстановления антигенов на последующих этапах работы иммунного алгоритма.

Антитело проходит отбор в том случае, если его средняя аффинность с антигенами выборки, определяемая по (6), удовлетворяет условию:

$$Af_{iAG} \geq 98\% \times CSel. \quad (7)$$

Оператор клонирования $clon(ab_1';...;ab_n')$ в рассматриваемом алгоритме использует приоритетное последовательное клонирование с конкурентно-целевым отбором клонов. Непосредственно процедура клонирования и отбора клонов может быть представлена в виде следующей последовательности операторов:

$$\begin{gathered} clon(ab_1;...;ab_n) = clon(ccreate(ab_1;...;ab_n), \\ cmut(ab_1';...;ab_n'), csel(ab_1'';...;ab_n'')), \\ (ab_1';...;ab_n') = create(ab_1;...;ab_n), \\ (ab_1'';...;ab_n'') = mut(ab_1';...;ab_n'), \end{gathered} \quad (8)$$

где $ccreate(ab_1;...;ab_n)$ – оператор создания клонов; $cmut(ab_1';...;ab_n')$ – оператор мутации клеток, а $csel(ab_1'';...;ab_n'')$ – оператор отбора клонов.

В [6, 7] клонирование выполняется одинаковым образом и заключается в том, что для каждого клонируемого антитела отводится некоторое количество клонов (от их общего числа), определяемое в зависимости от его параметров (средней аффинности с антигенами). Особенностью такой организации работы оператора клонирования является создание клонов одновременно для всей популяции антител. При таком распределении клонов каждый объект получает небольшое количество клонов, что уменьшает вероятность восстановления исходных антигенов. Это приводит к тому, что на восстановление антигенов требуется большое количество поколений антител и, следовательно, большие затраты времени. В предлагаемом алгоритме реализован метод приоритетного последовательного клонирования, суть которого заключается в том, что для каждого клонируемого объекта независимо от его параметров выделяется максимально возможное количество клонов. Создание, мутация и отбор клонов происходит последовательно для каждого клонируемого антитела, а не для всей популяции одновременно, как в [6, 7]. Благодаря этому растет площадь разброса клеток и увеличивается вероятность восстановления антигенов, однако при этом большое значение приобретает процедура отбора клонов. Суть предлагаемого метода конкурентно-целевого отбора заключается в том, что для каждого клонируемого объекта (антитела) определяется расстояние поиска целевых антигенов r:

$$r = k \times \sqrt{\frac{width \times height}{n}}, \quad (9)$$

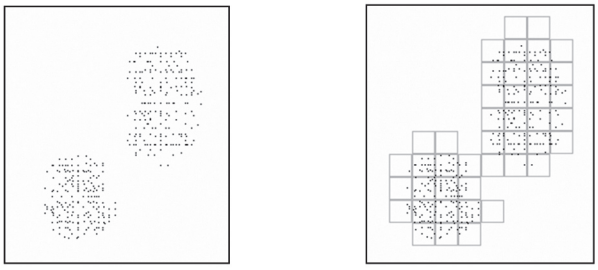
где *width* – ширина; *height* – высота области кластеризации; n – количество антигенов; k – коэффициент увеличения области поиска. Для всех клонов объекта после их создания и мутации определяются аффинности с целевыми антигенами, находящимися в области поиска r, после чего для каждого клона по максимальному значению аффинности определяется целевой антиген. Отбор клонов осуществляется за счет конкуренции клеток по целевому признаку, то есть из всего множества клонов, имеющих общий целевой антиген, отбирается тот, аффинность которого к данному антигену больше. При такой организации клонирования может возникнуть ситуация, при которой объектам для клонирования может быть недостаточно клонов. Для решения данной проблемы используются приоритеты. Это выражается в том, что объекты, которым не хватило клонов, на следующем шаге

будут обладать повышенным приоритетом и будут клонироваться в первую очередь.

После выполнения операторов отбора, клонирования и старения популяции антител остается некоторое количество неиспользованных антител. Для соблюдения равенства популяций антигенов и антител необходимо производить дополнительный разброс. В работе оператора дополнительного разброса предлагается формирование ограниченных областей разброса. В иммунных алгоритмах [6, 7] дополнительный разброс производится случайным образом. Вероятность восстановления антигенов при такой организации разброса невелика. В предлагаемом алгоритме кластеризации при организации дополнительного разброса использовался подход, ориентированный на формирование ограниченных областей с использованием контрольных точек.

Перед началом работы оператора восстановления выборки все пространство обработки данных разбивается на области дополнительного разброса со стороной r (9). После этого в каждой полученной области дополнительного разброса определяются контрольные точки, значения аффинностей в которых будут характеризовать область. Таким образом, на каждую область отводится всего пять точек, в которых при дополнительном разбросе устанавливаются антитела. Затем по аффинностям контрольных точек с восстанавливаемой популяцией антигенов вычисляется средняя аффинность всей области дополнительного разброса. В том случае, если аффинность области дополнительного разброса не удовлетворяет условию (7), область исключается из списка областей дополнительного разброса и не учитывается при разбросе антител. Таким образом, устанавливаются контуры больших областей дополнительного разброса, располагающихся на скоплении антигенов, а удаленные от таких скоплений области не будут учитываться.

На рис. 1 приведен пример формирования областей дополнительного разброса в местах скопления антигенов предлагаемым методом.



а *б*

Рис. 1. Организация дополнительного разброса:
а – исходное расположение антигенов;
б – области дополнительного разброса

Определение исходных кластеров – важнейшая задача, решаемая при кластеризации данных. В рассматриваемом алгоритме предлагается новый подход к решению данной задачи. Процесс формирования кластеров можно представить в виде следующей последовательности шагов:

Шаг 1. Из популяции $AB(ab_1;...;ab_n)$ с неопределенными кластерами, определяется антитело с максимальной аффинностью Af_{iAG}, которое будет являться условным центром нового кластера.

Шаг 2. Из популяции $AB(ab_1;...;ab_n)$ с неопределенными кластерами, отбираются объекты, расположенные на расстоянии d (10) от центра кластера:

$$d = Af_{iAG} \times \sqrt{\frac{width \times height}{n}}, \qquad (10)$$

если таковых нет – кластер удаляется.

Шаг 3. Для каждого антитела, входящего в кластер, определяется расстояние d (10) и осуществляется процесс поиска соседних антител (происходит расширение кластера).

Шаг 4. Возвращение к Шагу 1 до тех пор, пока не останется некластеризованных антител, либо дальнейшее формирование новых кластеров будет невозможно.

Кластеризация антител происходит после восстановления большинства антигенов (то есть достижения некоторого заданного порога восстановления клеток).

Кластеризация популяции исходных антигенов производится по результатам кластеризации последней популяции антител. Для каждого невосстановленного антигена вычисляются аффинности со всеми существующими кластерами, после чего оно относится к ближайшему кластеру, то есть к кластеру, аффинность по отношению к центру которого выше.

Алгоритм кластеризации объектов можно представить в виде следующей последовательности:

1. Формирование начальной популяции антител.
2. Определение критерия отбора антител.
3. Определение областей дополнительного разброса.
4. Отбор антител по условию (7).
5. Цикл восстановления и кластеризации антител:

5.1. Отбор антител, полученных в результате дополнительного разброса.

5.2. Последовательное клонирование антител по приоритету.

5.3. Удаление клонированных антител, которые не восстановили антигены.

5.6. Дополнительный разброс антител.

5.7. Проверка критерия восстановления антител. В случае его достижения переход к п.5.8., иначе возвращение к п. 5.1.

5.8. Определение параметров кластеров антител.

5.8.1. Определение центра кластера.

5.8.2. Определение исходных границ кластера.

5.8.3. Расширение кластера.

5.8.4. Возврат к п. 5.8.1. до тех пор, пока возможно формирование новых кластеров.

6. Формирование кластеров антигенов по результатам кластеризации последней популяции антител.

В результате работы будет получено множество кластеров $CL(cl_1;...;cl_m)$, сформированных из популяции исходных антигенов $AG(ag_1;...;ag_n)$.

3. Результаты экспериментальных исследований

Тестирование иммунного алгоритма кластеризации производилось на изображении размером 270×240 точек. Выборка состояла из 1000 антигенов, сформированных случайным образом в четырех областях разброса.

За счет последовательной организации клонирования антитела уже на первой итерации восстановили более 25% антигенов. В данном примере порог восстановления был принят равным 80%. Так, на 5-й итерации алгоритма начался процесс кластеризации антител. Для восстановления 99,6% антигенов потребовалось 7 итераций. В результате кластеризации последней популяции антител и определения кластеров антигенов были сформированы 4 кластера (рис. 2).

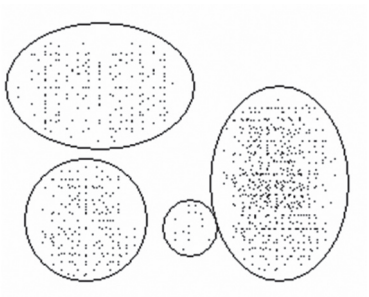


Рис. 2. Результаты кластеризации

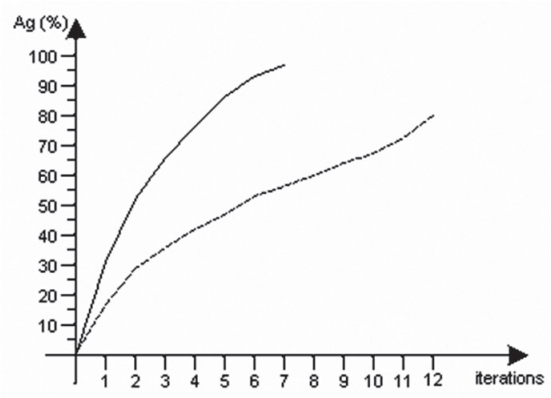


Рис. 3. Графики восстановления антигенов с использованием различных алгоритмов

На рис. 3 приведены графики восстановления антигенов с помощью предлагаемого алгоритма (сплошная линия) и алгоритма, приведенного в [6] (пунктирная линия). Следует отметить, что наибольшее количество случаев восстановления антигенов выборки приходится на первую итерацию. Затем эффективность восстановления клеток снижается от итерации к итерации с 20% на второй – до 8,5% на седьмой. Это обусловливается уменьшением количества возможных клонов на каждой последующей итерации.

Как видно по результатам тестирования, предлагаемый алгоритм практически полностью восстанавливает обучающую выборку за небольшое количество итераций, в то время как в соответствии с алгоритмом [6] для восстановления исходной популяции антигенов потребовалось в два раза больше поколений.

Выводы

В работе исследовались иммунные операторы, используемые для кластеризации объектов. В работе оператора клонирования было предложено использование приоритетного последовательного клонирования с конкурентно-целевым отбором клонов, что значительно ускорило процесс восстановления исходных антигенов. Для обеспечения работы оператора отбора антител было предложено использование критерия средней аффинности, что упростило процедуру отбора и повысило ее точность. В операторе дополнительного разброса антител был предложен метод разброса в определяемых ограниченных областях, что увеличило точность дополнительного разброса и ускорило процесс восстановления антигенов.

При формировании кластеров объектов был предложен новый подход к определению центров кластеров и их границ, позволяющий ускорить кластеризацию объектов.

Предложенный алгоритм кластеризации, несмотря на некоторую сложность в реализации, обеспечивает быстрое восстановление антигенов и распределение их по кластерам.

Список литературы: **1.** *A.K. Jain,* M.N. Murty, P.J. Flynn – «Data Clustering: A Review», ACM Computing Surveys, 1999, P. 265-323. **2.** *G. Gan,* C. Ma, J. Wu – «Data Clustering: Theory, Algorithms, and Applications», ASA-SIAM, Alexandria, 2007, 488 p. **3.** *W. Pedrycz,* «Knowledge-Based Clustering», Wiley&Sons, 2005, 377 p. **4.** *Искусственные* иммунные системы и их применение [Текст] / под ред. Д. Дасгупты : пер. с англ. А.А. Романюхи. – М.: ФИЗМАТЛИТ, 2006. – 344 с. **5.** *J. Timmis,* T. Knight, L.N. de Castro, E. Hart. «An Overview of Artificial Immune Systems»., Natural Computation, Springer, 2004, P. 55-86. **6.** *Lei Jia,* Licai Yang, Qingjie Kong, Shu Lin. «Study of Artificial Immune Clustering Algorithm and Its Applications to Urban Traffic Control», International Journal of Information Technology, Vol.12, No.3, 2006, 9 p. **7.** *Литвиненко, В.И.* Гибридная иммунная сеть для решения задач структурной идентификации [Текст]/ В.И. Литвиненко, П.И. Бидюк, А.А. Фефелов, И.В. Баклан // Нейронные сети. – № 1 – 2005. – С. 143-155.

Поступила в редколлегию 22.03.2010 г.

УДК 004.89

Дослідження імунних операторів в задачі кластеризації об'єктів / М.М. Корабльов, О.О. Фомічов // Біоніка інтелекту: наук.-техн. журнал. – 2010. – № 1 (72). – С. 70–74.

У статті досліджується робота імунних операторів, що використовуються при кластерізації об'єктів. При відновленні популяції антигенів використовується пріоритетне послідовне клонування з конкурентно-цільовим відбором клонів. В роботі оператора додаткового розкиду запропоновано метод розкиду об'єктів у визначених обмежених областях. Для організації відбору антитіл та клонів використається загальний критерій, що використовується також і при формуванні кластерів.

Іл. 3. Бібліогр.: 7 найм.

UDC 004.89

Study of immune operators in problems of clustering objects / N.M. Korabljov, A.A. Fomichev // Bionics of Intelligence: Sci. Mag. – 2010. – № 1 (72). – P. 70–74.

The paper investigates the work of the immune operators used in the clustering of objects. When restoring a population of antigen used priority consistent with the cloning competitive target selection of clones. In this paper, the operator further spread of a method for scatter objects in defined limited areas. To organize the selection of antibodies and clones used the general criterion, also used in the formation of clusters.

Fig. 3. Ref .: 7 items.