УДК 004.89

Н.М. Кораблёв, А.А. Фомичёв

Харьковский национальный университет радиоэлектроники, Харьков

КЛАСТЕРИЗАЦИЯ ДАННЫХ НА ОСНОВЕ ИСКУССТВЕННЫХ ИММУННЫХ СИСТЕМ

В работе рассматривается алгоритм кластеризации данных, работающий по принципам искусственной иммунной системы. Для решения задачи кластеризации предлагается использование приоритетного последовательного клонирования с конкурентно-целевым отбором клонов, критериев отбора клеток и определения кластеров. Данные подходы предлагаются впервые для использования в иммунных алгоритмах, работающих на принципах конкурентного отбора клеток. Экспериментальным путем было установлено, что использование предложенных подходов значительно ускоряет скорость восстановления выборки исходных антиген и кластеризации

Ключевые слова: *аффинность, критерий отбора антител, приоритетное последовательное клонирование, конкурентно-целевой отбор, дополнительный разброс, коэффициент увеличения области поиска.*

Введение

Основной задачей кластеризации данных является формирование кластеров, характеризующихся рядом признаков [1]. Входными параметрами для такого процесса является некоторое множество данных, подлежащих кластеризации. Для решения данной задачи можно использовать существующие подходы: от иерархических и k-means-алгоритмов, до алгоритмов, основанных на принципах работы нейронной сетей [1, 2]. Эти алгоритмы отличаются друг от друга, как по сложности реализации, так и по скорости работы и точности определения кластеров.

Задача кластеризации может быть решена на основе искусственных иммунных систем (ИИС) [3], в основе работы которых лежат принципы функционирования иммунной системы позвоночных, и, в частности, человека. ИИС обладают большим потенциалом, что обуславливает их использование в задачах обработки информации, таких как: кластеризация, классификация, распознавание образов, идентификация и др. Спецификой иммунных алгоритмов кластеризации является их способность восстанавливать исходные данные, которые представляются в виде популяции антиген, путем клонирования и отбора случайным образом сформированной популяции антител. В настоящее время существуют ИИС-алгоритмы кластеризации [4, 5], которые в решении основных задач (отбор, клонирование, определение кластеров) используют одни и те же подходы. Данные алгоритмы достаточно просты в реализации и эффективно кластеризуют данные, однако, на некоторых этапах их работы производятся избыточные вычислительные операции, результатом чего является снижение скорости их работы. В [4, 5] на этапах отбора антител и их клонов для каждой клетки, проходящей отбор, производится большое количество избыточных вычислений. Кроме того, в [4] отбор клеток производится не самым эффективным образом, а в [5] используется большое количество коэффициентов, усложняющих реализацию алгоритма и замедляющих его работу. Как в [4], так и в [5] решение задачи клонирования и отбора клеток приводит к значительным затратам времени выполнения. В этих

алгоритмах клонирование производится параллельно, при этом клонируемые клетки конкурируют между собой за количество клонов. Такой способ организации клонирования не обеспечивает максимального покрытия антиген, что увеличивает количество шагов в цикле восстановления антиген.

В предлагаемом алгоритме реализуются новые подходы к клонированию и отбору клонов и антител. Для отбора антител предлагается использование критерия, с помощью которого останутся все антитела, расположенные поблизости от скоплений антиген, а не какая-либо их часть, как в [5]. Кроме того, предполагается последовательное клонирование, при котором каждое антитело создает максимально возможное количество клонов, в результате чего покрывается (восстанавливается) большее количество антиген. Для отбора клонов антител используется метод конкурентно-целевого отбора. Все это сокращает избыточные вычисления и способствует ускорению работы алгоритма.

Постановка задачи

В ИИС-алгоритмах множество исходных данных, подлежащих кластеризации, представляется набором антиген $AG(ag_1;...;ag_n)$, каждый из которых обладает рядом признаков. Для восстановления и кластеризации входных данных случайным образом формируется исходная популяция антител $AB(ab_1;...;ab_n)$. Важной особенностью ИИС-алгоритмов является количественное равенство популяций клеток, т.е. количество антител должно быть равно количеству антиген, и это равенство должно соблюдаться на всех итерациях работы алгоритма. Кроме того, прямое использование параметров антиген возможно только при проведении процедуры отбора антител.

В общем виде ИИС-алгоритм кластеризации может быть представлен следующим образом:

$$\begin{gathered} clust(ag_1;...;ag_n) = clust(reconstr(sel(ab_1;...;ab_n), \\ clon(ab_1';...;ab_n'), ageing(ab_1'';...;ab_n''), \\ dispers(ab_1;...;ab_n))) \\ (ab_1';...;ab_n') = sel(ab_1;...;ab_n) \\ (ab_1'';...;ab_n'') = clon(ab_1';...;ab_n'), \end{gathered} \quad (1)$$

где $reconstr(...)$ – функция восстановления клеток; $sel(ab_1;...;ab_n)$ – функция отбора антител; $clon(ab_1';...;ab_n')$ – функция клонирования, мутации и отбора клонов для популяции антител; $ageing(ab_1'';...;ab_n'')$ – функция старения клеток; $dispers(ab_1;...;ab_n)$ – функция дополнительного разброса клеток. Решение каждой из этих задач существенно влияет на точность и скорость работы всего алгоритма кластеризации.

В работе иммунных алгоритмов процесс отбора популяции антител $sel(ab_1;...;ab_n)$ играет очень важную роль. Антитела, не удовлетворяющие условиям отбора, удаляются, а антитела, прошедшие отбор, клонируются. При осуществлении отбора клеток используется понятие аффинности как степени их близости [3-5]:

$$Af_{ij} = 1 + d_{ij}^{-1}. \quad (2)$$

Здесь Af_{ij} – аффинность между i и j клетками, а d_{ij} – евклидово расстояние между ними. Таким образом, критерий аффинности является основной определяющей мерой близости клеток.

Необходимо разработать ИИС-алгоритм кластеризации, функционирующий в соответствии с (1), использующий в качестве основной меры (2) и осуществляющий формирование кластеров из исходной выборки антиген.

Алгоритм кластеризации

Предлагаемый ИИС-алгоритм кластеризации отличается от существующих аналогов в решении следующих задач: отбор антител; клонирование и отбор клонов; осуществление дополнительного разброса; определение исходных кластеров.

Процедура отбора антител в ИИС-алгоритмах начинается с определения для каждого антитела из всей популяции $AB(ab_1;...;ab_n)$ средней аффинности Af_{iAG} со всеми антигенами $AG(ag_1;...;ag_n)$:

$$Af_{iAG} = \sum_{j=1}^{n} Af_{iag_j} \Big/ n, \quad (3)$$

где Af_{iag_j} – аффинность антитела ab_i и антигена ag_j. Непосредственно отбор клеток осуществляется с использованием критерия отбора антител $CSel$ как средней аффинности антиген:

$$CSel = \sum_{i=1}^{n} AF_i \Big/ n, \quad (4)$$

где AFi – средняя аффинность ag_i со всеми антигенами $AG(ag_1;...;ag_n)$.

Антитело проходит отбор в двух случаях:

– средняя аффинность Af_{iAG} антитела ab_i удовлетворяет условию (5):

$$Af_{iAG} \geq 98\% \times CSel; \quad (5)$$

– среднюю аффинность Af_{iAG} невозможно вычислить.

В первом случае в условии (5) для отбора клеток используется некоторая доля аффинности от критерия отбора $CSel$. Это объясняется тем, что при требовании полного совпадения средней аффинности Af_{iAG} с критерием $CSel$ количество проходящих отбор антител будет незначительным, либо та-

ких антител вообще не будет. Это может привести к удалению всей популяции и повторению процедуры генерации и отбора антител. Во втором случае невозможность вычисления аффинности Af_{iAG} свидетельствует о том, что антитело ab_i совпало с каким-либо антигеном популяции $AG(ag_1;...;ag_n)$, т.е. произошел процесс восстановления клетки на первой популяции. В результате процедуры отбора антител, организованного описанным способом, остаются все антитела, расположенные на достаточно небольшом расстоянии от скопления антиген, а не их часть, как в [4, 5].

Формально процедура клонирования и отбора описывается следующим образом:

$$clon(ab_1';...;ab_n') = funct(create(cl_1;...;cl_m), mut(cl_1;...;cl_m), sel(cl_1;...;cl_m), add((cl_1';...;cl_z')(ab_1;....;ab_n))), \quad (6)$$

где $create(cl_1;...;cl_m)$ – функция создания клонов; $mut(cl_1;...;cl_m)$ – функция мутации клонов; $sel(cl_1;...;cl_m)$ – функция отбора клонов, а $add((cl_1';...;cl_z')(ab_1;....;ab_n))$ – функция добавления отобранных клонов $(cl_1';...;cl_z')$ в множество антител $AB(ab_1;...;ab_n)$, при $m = n_{ag} - n_{ab}$, $z < m$.

Клонирование и отбор клонов осуществляется последовательно с использованием приоритетов. При последовательном клонировании каждое клонируемое антитело имеет максимально возможное количество клонов, за счет чего повышается вероятность покрытия (восстановления) большего количества антиген. Однако при этом особое значение приобретает процедура отбора клонов. Это объясняется тем, что при недостаточно эффективной организации процедуры отбора клонов возрастает вероятность события, при котором для клонирования некоторого количества антител не будет хватать клонов. Вследствие этого популяция антител на итерациях алгоритма не будет клонирована полностью. Для решения данной задачи используется приоритетное клонирование и конкурентно-целевой отбор клонов. Суть приоритетного клонирования заключается в том, что не каждой i итерации алгоритма в первую очередь произойдет клонирование антител, которые не были клонированы на $i-1$ итерации.

Для повышения эффективности процедуры отбора и ее ускорения предлагается использовать конкурентно-целевой отбор клонов. Суть последовательного клонирования заключается в том, что при отборе клонов антител используются не все антигены, а лишь та их часть, которая расположена на расстоянии r от клонируемого антитела:

$$r = k \times \sqrt{\frac{width \times height}{n}}, \quad (7)$$

где $width$ – ширина; $height$ – высота области кластеризации; n – количество антиген; k – коэффициент увеличения области поиска. Антигены, находящиеся в сформированной области, формируют массив возможных целей $TG\{ag_k;...;ag_f\}$ для клонов антитела. При этом отбираются только не восстановленные антигены. Кроме антиген в данной области учитываются и присутствующие антитела для предотвращения ситуации совпадения с ними клонов. После клонирования и мутации начинается процедура конкурентного отбора. Для каждого клона определяются аффинности с возможными целями $TG\{ag_k;...;ag_f\}$. Затем из множества возможных целей определяется антиген с наивысшей аффинностью по отношению к клону. Во время процедуры конкурентного отбора из каждой пары клонов с идентичными целями удаляется тот, аффинность с целью которого меньше.

Дополнительный разброс антител производится в том случае, когда количество антител меньше количества антиген. Организация дополнительного разброса также имеет множество решений. Обычно дополнительный разброс производится случайным образом по всей поверхности выборки. Такой метод прост, но не дает хороших результатов, поскольку вероятность попадания антитела в результате дополнительного разброса в необходимую область с антигенами достаточно невелика. Для обеспечения наиболее точного разброса предлагается использовать разброс клеток в заранее сформированных областях. Для этого в соответствии с (7) всё пространство выборки разделяется условно на квадраты со стороной r. Разброс в таких квадратах будет производиться строго определённым образом – всего разбрасывается по 5 клеток. Это производится для того, чтобы при использовании минимального количества клеток на область вероятность их близкого расположения с антигенами, находящимися в области была достаточной. Первая точка устанавливается в центре области, остальные – на расстоянии $k = (1/2) \times r$ от центральной точки по осям координат во всех направлениях.

Определение исходных кластеров – важнейшая задача, решаемая при кластеризации данных. В рассматриваемом алгоритме предлагается новое решение данной задачи. После определения для каждого антитела с неустановленным кластером средней аффинности со всеми антигенами Af_{iAG} определяется антитело с максимальной аффинностью. Данное антитело считается исходным центром нового кластера. Затем определяются антитела, расположенные на расстоянии d по осям от координат центра кластера:

$$d = Af_{iAG} \times \sqrt{\frac{width \times height}{n}}. \quad (8)$$

Эти антитела входят в сформированный кластер. Затем для каждого антитела, входящего в кластер, опре-

деляется расстояние d и осуществляется процесс поиска соседних антител. Так происходит расширение кластера. Этот процесс производится до тех пор, пока вся популяция антител не будет кластеризована.

Кластеризация антител начинается после восстановления большинства антиген (т.е. достижения некоторого заданного порога восстановления клеток). Непосредственное формирование кластеров антител производится по описанному ранее алгоритму. Окончательная кластеризация антиген производится по результатам кластеризации последней популяции антител. Для каждого не восстановленного антигена вычисляются аффинности со всеми существующими кластерами, после чего оно относится к ближайшему кластеру, т.е. к кластеру, аффинность по отношению к центру которого, выше.

Алгоритм кластеризации можно представить в виде следующей последовательности шагов:

1. Формирование начальной популяции антител.
2. Определение восстанавливающих антител.
3. Определение областей дополнительного разброса.
4. Отбор антител.
5. В случае удаления всех антител возвращение к шагу 1.
6. Цикл восстановления и кластеризации антител:

6.1. Проверка возможности проведения итерации.

6.2. Отбор антител полученных в результате дополнительного разброса на предыдущей итерации

6.3. Удаление клонированных не восстанавливающих антител.

6.4. Последовательное клонирование антител по приоритету.

6.4.1. Определение множества целевых антиген для клонов.

6.4.2. Определение антител в области клонирования.

6.4.3. Создание и мутация клонов антитела.

6.4.4. Определение восстанавливающих клонов.

6.4.5. Определение целей для не восстанавливающих клонов.

6.4.6. Конкурентный отбор клонов.

6.4.7. Перенос оставшихся клонов в массив антител.

6.5. Установка флага приоритетного клонирования для всех «старых» антител, клонирование которых не произошло.

6.6. Осуществление дополнительного разброса.

6.8. Определение параметров кластеров антител на итерации.

7. Формирование кластеров антиген.
8. Конец.

Результатом является кластеризованная популяция антиген.

Экспериментальные исследованияя

Тестирование предлагаемого алгоритма кластеризации производилось на изображении размером 290×220 точек (рис. 1). Выборка состояла из 800 антиген, сформированных случайным образом в четырех областях разброса. После формирования выборки, по параметрам антиген, был определен критерий отбора клеток.

После формирования первой популяции антител и определения восстанавливающих антител начался процесс отбора клеток. В результате отбора из исходной популяции остались лишь антитела наиболее приближенные к скоплениям антиген и восстанавливающие клетки.

После отбора начался процесс восстановления антиген за счет клонирования и мутации поколений антител. За счет последовательной организации клонирования антитела уже на первой итерации покрыли большие области клонами и восстановили до 30% антиген. На рис. 2 приведен результат после первой итерации алгоритма.

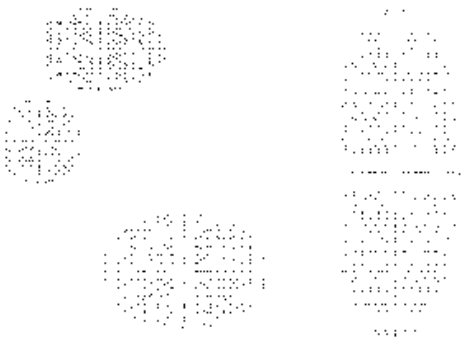

Рис. 1. Исходная выборка антиген

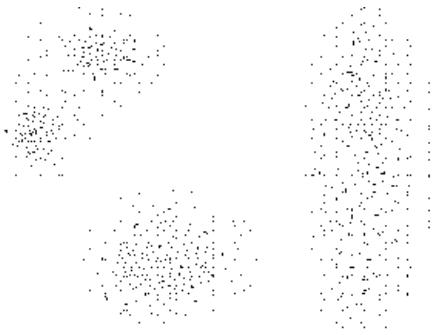

Рис. 2. Результат работы алгоритма после первой итерации

Коэффициент увеличения области поиска k в данном примере был принят равным $1{,}8$. Это объясняется тем, что при большем значении k количество целевых антиген увеличивается, что ведет к росту вычислений. При меньшем значении k количество целевых антиген сокращается, что может привести к ситуации, когда ни одной цели для клонов не будет найдено.

Большое значение в работе алгоритма приобретает порог восстановления клеток, после которого начинается определение кластеров. Это объясняется тем, что процесс формирования кластеров для популяции антител связан с большим количеством вычислений, что может существенно увеличить время выполнения алгоритма. Поэтому в данном примере

порог восстановления был установлен равным 80% . На 5-ой итерации алгоритма начался процесс кластеризации антител. Для восстановления 99,7% антиген алгоритму потребовалось 7 итераций.

В результате кластеризации последней популяции антител и определения кластеров антиген были сформированы 4 кластера (рис. 3). На рис. 4 представлен график восстановления антиген на итерациях алгоритма. Следует отметить, что наибольшее количество случаев восстановления антиген выборки приходится на первую итерацию. Затем эффективность восстановления клеток снижается от итерации к итерации: с 20% на второй – до 8,5% на седьмой. Это обуславливается значительным уменьшением количества возможных клонов на каждой последующей итерации.

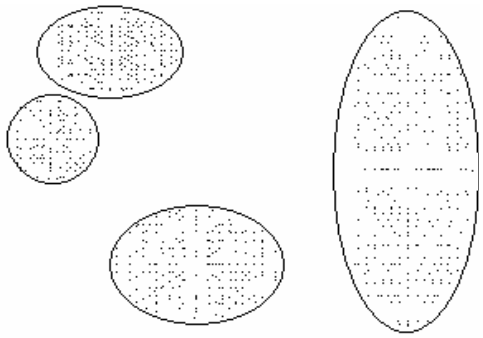

Рис. 3. Кластера антиген сформированные в ходе работы алгоритма

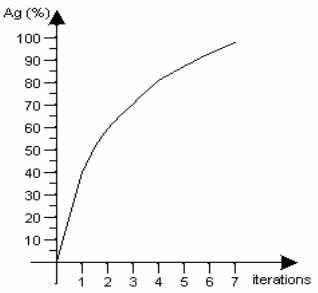

Рис. 4. График восстановления антиген на итерациях алгоритма

Как видно по результатам тестирования, данный алгоритм практически полностью восстанавливает обучающую выборку за небольшое количество итераций, и обладает хорошей точностью в определении кластеров. Кроме того, он также отличается высокой скоростью работы благодаря организации конкурентно-целевого отбора клонов и увеличению точности дополнительного разброса клеток.

Выводя

Разработан ИИС-алгоритм кластеризации данных, в котором используются новые подходы для решения основных задач. Для отбора клеток используется единый критерий отбора, что значительно повышает эффективность всего алгоритма. Использование приоритетного последовательного клонирования ускоряет процесс восстановления исходной выборки антиген, что позволяет произвести более точную кластеризацию клеток.

Конкурентно-целевой отбор клонов в ограниченных областях поиска является принципиально новым подходом к решению задачи отбора клонов. Благодаря ограничению области поиска целей значительно сокращается объем вычислений. Использование конкурентно-целевого отбора клеток позволяет провести наиболее эффективный отбор клонов при минимальных временных затратах.

В описанном алгоритме предложены новые подходы для решения задачи дополнительного разброса клеток и определения исходных кластеров антител. Использование перечисленных подходов повышает скорость работы алгоритма при восстановлении исходного множества антиген (выборки) и определения их кластеров. Предложенный алгоритм кластеризации, несмотря на некоторую сложность в реализации, обеспечивает быстрое восстановление антиген и распределение их по кластерам. При незначительной модификации данный алгоритм может использоваться для решения задач классификации, идентификации и распознавания образов.

Списокяитературыя

1. *Jain A.K. Data Clustering: A Review / A.K. Jain, M.N. Murty, P.J. Flynn. – 220 p.*

2. *Kogan J. Clustering Large and High Dimensional Data / J. Kogan, C. Nikolas, M. Teboulle. – 188 p.*

3. *Искусственные иммунные системы и их применение / Под ред. Д. Дасгупты, пер. с англ. А.А. Романюхи. – М.: ФИЗМАТЛИТ, 2006. – 344 с.*

4. *Study of Artificial Immune Clustering Algorithm and Its Applications to Urban Traffic Control / Lei Jia, Licai Yang, Qingjie Kong, Shu Lin // International Journal of Information Technology. – 2006. – Vol. 12, No. 3. – P. 125-131.*

5. *Гибридная иммунная сеть для решения задач структурной идентификации / В.И. Литвиненко, П.И. Бидюк, А.А. Фефелов, И.В. Баклан. – 320 с.*

Поступила в редколлегию 17.06.2009

Рецензент: д-р техн. наук, проф. С.Г. Удовенко, Харьковский национальный университет радиоэлектроники, Харьков.

КЛАСТЕРИЗАЦІсяДАНИХяНАяОСНОВІяШТУЧНИХяІМУННИХяСИСТЕМя

М.М. Корабльов, О.О. Фомічов

В роботі розглядається алгоритм кластеризації даних, що працює за принципами штучної імунної системи. Для вирішення задачі кластеризації пропонується використання пріоритетного послідовного клонування із конкурентно-цільовим відбором клонів, критеріїв відбору клітин й визначення кластерів. Ці підходи вперше пропонуються для використання в імунних алгоритмах, що працюють на принципах конкурентного відбору клітин. Експериментальним шляхом було встановлено, що використання підходів що пропонуються значно прискорює швидкість відновлення вибірки антиген та кластеризації.

Ключові слова: *аффінність, критерій відбору антитіл, пріоритетне послідовне клонування, конкурентно-цільовий відбір, додатковий розкид, коефіцієнт збільшення ділянки пошуку.*

DATA CLUSTERING USING ARTIFICIAL IMMUNE SYSTEMS

N.M. Korablev, A.A. Fomichov

The given work is devoted to the designing of the data clustering algorithm based on artificial immune system approach. It proposed to use priority sequential cloning with competitive targeting cloning selection and some selection and clustering criteria for resolve clustering problem. Those methods proposed in this work at first. By the experiment way it was detected, that using of this methods makes faster speed of antigens sampling remaking and clustering process

Keywords: *affinity, antibodies selection criterion, priority sequential cloning, competitive targeting cloning selection, additional dispersion, searching space growth criterion.*