

УДК 004.89

Н.М. Кораблев¹, А.А. Фомичев²¹ХНУРЭ, г.Харьков, Украина, korablev@kture.kharkov.ua²ХНУРЭ, г.Харьков, Украина, alexandros_1985@mail.ru

КЛАСТЕРИЗАЦИЯ ДАННЫХ МЕТОДОМ K-MEANS С ИСПОЛЬЗОВАНИЕМ ИСКУССТВЕННЫХ ИММУННЫХ СИСТЕМ

В данной работе рассматривается алгоритм кластеризации данных методом k-means, функционирующий на основе использования искусственных иммунных систем. При определении центров k кластеров используется значение средней пороговой аффинности объектов. Для повышения скорости работы алгоритма кластеризации используются принципы клонирования граничных объектов и клонирования центров (центроидов).

ИСКУССТВЕННЫЕ ИММУННЫЕ СИСТЕМЫ, ПОСЛЕДОВАТЕЛЬНОЕ КЛОНИРОВАНИЕ, ПОРОГОВАЯ АФФИННОСТЬ, УРОВЕНЬ СТИМУЛЯЦИИ

Введение

В настоящее время проблема кластеризации данных является одной из наиболее актуальных в области информационных технологий. Среди большого количества существующих методов кластеризации данных одним из часто используемых является метод k-means [1, 4]. Данный метод отличается тем, что количество кластеров изначально известно и определяется числом k . Характерными особенностями данного метода являются его простота реализации и модификации. Однако методу k-means присущи некоторые недостатки – проблема сходимости, и, как следствие, невозможность определения времени, необходимого на кластеризацию данных.

Для повышения скорости работы алгоритма без потери точности при классификации методом k-means предлагается использование биологических принципов организации вычислений, основанных на принципах работы искусственных иммунных систем (ИИС) [2, 3, 5]. При этом большое внимание уделяется определению исходных центров k кластеров, а также работе операторов клонирования и мутации клонов. Характерной особенностью иммунной модификации метода k-means является ограничение количества объектов, которые будут клонированы, на основе критерия пороговой аффинности, что значительно ускоряет работу алгоритма.

1. Постановка задачи

Дано множество объектов n , каждый из которых описывается набором признаков, и количество искомым кластеров k . Исходные объекты представляются популяцией антиген $AG(ag_1; \dots; ag_n)$. В качестве меры близости между объектами используется критерий аффинности Af_{ij} [2, 3, 5]:

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

где Af_{ij} – аффинность между i и j объектами, а d_{ij} – евклидово расстояние между их признаками.

Необходимо разработать алгоритм кластеризации объектов, функционирующий на основе метода k-means и иммунных принципов группировки, использующий в качестве основной меры близости объектов критерий (1), позволяющий определять центры k кластеров и принадлежность к ним исходных объектов.

2. Алгоритм кластеризации

Работу алгоритма k-means можно условно разделить на четыре основных этапа [1, 4]:

1. Определение k центров кластеров.
2. Определение принадлежности объектов кластерам.
3. Определение центроидов k кластеров.
4. Сравнение центров и центроидов кластеров.

Определение центров кластеров может производиться как случайным образом, так и на основе некоторых критериев (например, по значению среднего евклидова расстояния с остальными объектами). Центроиды являются центрами кластеров, они могут быть кластеризуемыми объектами и определяются по средним значениям признаков объектов, входящих в данный кластер. На последнем этапе производится сравнение центров (центроидов), полученных на предыдущей итерации алгоритма, и центроидов, полученных после переопределения принадлежности объектов кластерам. В случае их совпадения алгоритм кластеризации завершает работу, иначе происходит возврат ко второму этапу, где будет определяться принадлежность объектов к кластерам на основании выделенных ранее центроидов.

Использование ИИС–принципов в организации работы k-means приводит к некоторому изменению в организации вычислений. Иммунная система оперирует с двумя типами клеток (лимфоцитов): Т-клетки, распознающие антигены и стимулирующие иммунный ответ, и В-клетки, нейтрализующие антигены. Антигенами $AG(ag_1; \dots; ag_n)$ являются все исходные объекты, В-клетки [3]

формируют популяцию $Vc(bc_1; \dots; bc_k)$ и являются кластерами, которые переопределяются в ходе работы алгоритма. В работе данного метода на начальном этапе используется популяция Т-клеток. Их основная задача заключается в определении критерия пороговой аффинности Aff_{thr} и стимуляции В-клеток.

Основная идея модифицированного k-means алгоритма при использовании ИИС состоит в том, что предлагаемая модель иммунной сети предполагает, что процесс поиска антигенов уже произведен и Т-клетки, вступившие во взаимодействие с антигенами, определили их признаки (априорная завершенность процесса распознавания). В результате распознавания популяция Т-клеток содержит признаки всех исходных антигенов, после чего для каждой Т-клетки определяется уровень стимуляции. После отбора k Т-клеток с наибольшим уровнем стимуляции формируется популяция В-клеток путем клонирования отобранных объектов. После этого популяция В-клеток клонируется и мутирует до тех пор, пока не будут выделены все В-клетки, каждая из которых в ходе клонирования и мутации станет соответствующей некоторому количеству входных антигенов.

Формально предлагаемый иммунный алгоритм k-means может быть представлен в виде следующей последовательности операторов:

$$\begin{aligned} kmAIS(ag_1; \dots; ag_n) = & stimTcell(t_1; \dots; t_n), \\ selBmem(b_1; \dots; b_k), & clonBcell(b'_1; \dots; b'_k), \\ & stimBcell(b'_1; \dots; b'_k), \\ matchBcell((b_1; \dots; b_k), & (b'_1; \dots; b'_k)), \end{aligned} \quad (2)$$

где $stimTcell(t_1; \dots; t_n)$ – оператор определения уровня стимуляции Т-клеток; $selBmem(b_1; \dots; b_k)$ – оператор выделения центров В-клеток (исходных центров кластеров); $clonBcell(b'_1; \dots; b'_k)$ – операторы клонирования, мутации и отбора В-клеток (формирование центроидов); $stimBcell(b'_1; \dots; b'_k)$ – оператор определения уровня стимуляции для полученных центроидов (определение принадлежности исходных объектов выделенным кластерам); $match((b_1; \dots; b_k), (b'_1; \dots; b'_k))$ – оператор сравнения клеток памяти $Bmem(b_1; \dots; b_k)$ и сформированных клеток $Bcell(b'_1; \dots; b'_k)$.

Следует отметить, что работа оператора $clonBcell(b'_1; \dots; b'_k)$ может происходить на основе следующих подходов:

1. Клонирование граничных антител в В-клетке.
2. Клонирование центра (ядра) В-клетки.

Граничные антитела В-клетки определяются с помощью критерия пороговой аффинности Aff_{thr} . Это происходит из предположения, что все антитела выделенной В-клетки соответствуют антигенам в ее области, т.е. её антитела по своим признакам идентичны вступившим с ней во взаимодействие

антигенам. После этого граничные антитела подвергаются клонированию и мутации. Принадлежность объекта (антигена) кластеру (В-клетке) определяется путем выбора максимального веса клонов объектов (антител), соответствующего одной из В-клеток (кластеров).

При клонировании всей В-клетки происходит мутация ее центра, представленного антителом с максимальным уровнем стимуляции ко всем антигенам, взаимодействующим с В-клеткой, либо её ядром (центроидом). После клонирования и мутации центра новое ядро (центроид) отбирается среди полученных клонов на основании конкурентного отбора, при этом ядром (центроидом) становится клон с наибольшим уровнем стимуляции ко всем антигенам В-клетки.

Для определения уровня стимуляции Т-клеток $stimTcell(t_1; \dots; t_n)$ производится вычисление их средних аффинностей aff_{iAG} со всеми антигенами популяции $AG(ag_1; \dots; ag_n)$:

$$aff_{iAG} = \frac{\sum_{j=1}^n aff_{ij}}{n}, \quad (3)$$

где aff_{ij} – аффинность Т-клетки t_i и антигена ag_j .

После этого на основании полученных уровней стимуляции производится определение значения пороговой аффинности Aff_{thr} :

$$Aff_{thr} = \frac{\sum_{i=1}^n aff_{iAG}}{n}. \quad (4)$$

Пороговая аффинность используется при определении центров В-клеток (центров кластеров) и определении граничных антител (граничных объектов кластера) при кластеризации.

Перед формированием популяции В-клеток по значениям aff_{iAG} определяются k наиболее стимулированных Т-клеток, аффинность между которыми не превышает значения Aff_{thr} . Эти клетки формируют популяцию В-клеток, каждая из которых изначально характеризуется только одним антителом (центром). Данное антитело становится клеткой памяти системы. После определения центра происходит рост В-клеток – антитела относятся к В-клеткам по максимальным значениям аффинности с её ядром (центральным объектом, центроидом).

Клонирование и мутация В-клеток $clonBcell(b'_1; \dots; b'_k)$ может происходить на основе указанных подходов. В первом случае (клонирование граничных антител) из антител, входящих в В-клетку для клонирования, выбираются антитела, аффинность которых к центру В-клетки удовлетворяет выражению:

$$aff_{ij} \leq \gamma \cdot Aff_{thr}, \quad (5)$$

где aff_{ij} – аффинность i -го антитела с j -ым центром В-клетки (центром кластера), а γ – коэффициент деформации (сжатия) клетки, устанавливаемый в процентах. Для каждого клонируемого антитела количество возможных клонов определяется следующим образом:

$$cl_i = int \parallel (k \cdot aff_{ij})^2 \parallel, \quad (6)$$

где $int \parallel$ – округляет значение до ближайшего большего целого, k – количество кластеров, а aff_{ij} – аффинность i -го антитела с центром j -ой В-клетки.

При мутации клонированных антител используется оператор гипермутации, в результате чего коэффициент мутации σ_i определяется как:

$$\sigma_i = \gamma \cdot (1 - aff_{ij}), \quad (7)$$

где aff_{ij} – аффинность i -го антитела с j -ым центром В-клетки, а γ – коэффициент деформации (сжатия) клетки, устанавливаемый в процентах.

После клонирования и мутации клоны антитела определяют аффинности со всеми k центрами В-клеток (центрами кластеров). Принадлежность антитела кластеру определяется по значению общей максимальной аффинности его клонов ядру (центру, центроиду) В-клетки. После этого происходит переопределение ядер (центроидов) В-клеток по средним аффинностям входящих в него антител.

Во втором случае (клонирование центра В-клетки) клонируется только ядро (центральный антиген), при этом количество клонов определяется как количество антител, содержащихся в данной В-клетке, а коэффициент мутации определяется как:

$$\sigma_i = \gamma \cdot aff_{iAG}, \quad (8)$$

где aff_{iAG} – аффинность центра i -й В-клетки с антигенами множества $AG(ag_1; \dots; ag_n)$.

После клонирования для каждого из клонов определяется средняя аффинность aff_{iAG} с антигенами $AG(ag_1; \dots; ag_n)$. В результате конкурентного отбора [5] остается только клон с наибольшим значением аффинности, который формирует новое ядро (центроид) мутировавшей В-клетки.

При определении уровня стимуляции полученных ядер (центроидов) В-клеток $stimBcell(b'_1; \dots; b'_k)$ для каждого нового ядра вычисляется аффинность aff_{iAG} , которая определяет уровень стимуляции В-клетки.

В работе оператора сравнения $match((b_1; \dots; b_k), (b'_1; \dots; b'_k))$ происходит определение аффинности между соответствующими парами центров В-клеток. Клонирование и мутация В-клетки прекращается, если выполняются следующие условия:

$$aff_{ij} \geq \frac{\gamma \cdot Aff_{thr}}{2}; \quad \sigma_i - \sigma_j \leq \frac{\gamma \cdot Aff_{thr}}{2}, \quad (9)$$

где γ – коэффициент деформации клетки; Aff_{thr} – пороговая аффинность; aff_{ij} – аффинность между соответствующей парой центров; σ_i и σ_j – уровни стимуляции антител.

В случае, если данное условие не выполняется, в памяти происходит замещение предыдущего центра В-клетки новым ядром (центроидом) и операции клонирования, стимуляции и сравнения повторяются.

Процесс кластеризации с помощью иммунного k -means алгоритма можно представить как следующую последовательность шагов:

1. Определение уровней стимуляции aff_{iAG} (3) для популяции Т-клеток и значения пороговой аффинности Aff_{thr} (4) для формируемой популяции В-клеток.

2. Определение k наиболее стимулированных Т-клеток, аффинность между которыми не превышает значения пороговой аффинности Aff_{thr} . Формирование исходной популяции В-клеток по результатам отбора Т-клеток.

3. Определение принадлежности свободных антител В-клеткам по их аффинностям с центрами В-клеток.

4. Клонирование и мутация В-клеток по одному из предложенных вариантов.

5. Определение принадлежности антител к мутировавшим В-клеткам по аффинностям с их центрами.

6. Сравнение центров В-клеток, расположенных в памяти, с центрами, выделенными в результате клонирования и мутации (8). В случае невыполнения условий (9) замещение клеток памяти выделенными центрами и переход к шагу 4, иначе переход к шагу 7.

7. Конец.

В результате работы будет получено множество k В-клеток (кластеров), содержащих антитела. Для антигенов принадлежность к кластерам (В-клеткам) определяется по принадлежности антител, вступивших с ними во взаимодействие.

3. Результаты экспериментальных исследований

Тестирование алгоритмов кластеризации производилось на нескольких наборах данных (табл. 1).

Таблица 1

Характеристики наборов данных

| Характеристики | Набор 1 | Набор 2 | Набор 3 |
|-----------------------------|---------|---------|---------|
| Количество объектов | 100 | 1000 | 10000 |
| Количество групп признаков | 2 | 5 | 10 |
| Размерность групп признаков | 3 | 6 | 10 |
| Кол. кластеров k | 3 | 8 | 14 |

На первом этапе работы алгоритмов для определения исходных центров кластеров по стан-

дартному методу k-means потребовалось меньше времени, чем его иммунным модификациям. Этап роста выделенных кластеров потребовал от всех типов алгоритмов приблизительно одинаковых затрат времени, но стандартный метод k-means снова показал лучший результат по сравнению с двумя его иммунными модификациями. Аналогичный результат наблюдался также и на этапе выделения новых центров (центроидов) k кластеров, где стандартному алгоритму k-means снова потребовалось меньше времени, в то время как для иммунных методов на данном этапе характерны наибольшие затраты времени, поскольку клонирование и мутация объектов сопряжены с выполнением большого количества вычислительных операций. При этом важной особенностью является наблюдаемое различие между временными затратами различных иммунных модификаций k-means: подход клонирования граничных антител при небольших наборах данных, характеризующихся малым количеством объектов, небольшим количеством групп признаков и их размерностью, требует больших временных затрат, чем подход клонирования центра кластера. Однако при повышении количества объектов в наборе данных, повышении количества групп признаков или повышении размерности при клонировании граничных объектов затраты времени меньше, чем при клонировании и мутации центра В-клеток. На сравнение исходных центров и выделенных центроидов алгоритмам потребовалось приблизительно одинаковое количество времени.

Несмотря на то, что на выполнение одного прохода цикла кластеризации стандартному алгоритму k-means потребовалось значительно меньше времени, чем иммунным аналогам, его общие затраты времени больше, т.к. для кластеризации ему потребовалось значительно большее количество проходов цикла, чем его иммунным модификациям (табл. 2).

Таблица 2

Результаты работы алгоритмов

| Алгоритмы | Набор 1 | | Набор 2 | | Набор 3 | |
|----------------------------|---------|-----|---------|-----|---------|-----|
| | С | Т | С | Т | С | Т |
| Стандартный k-means | 13 | 100 | 154 | 100 | 1816 | 100 |
| Иммунный k-means 1-го типа | 9 | 96 | 96 | 87 | 1032 | 50 |
| Иммунный k-means 2-го типа | 7 | 88 | 93 | 83 | 1029 | 57 |

В табл. 2 представлены результаты работы алгоритмов кластеризации на различных наборах данных, где С – количество проходов цикла кластеризации для выделенного набора данных, Т – общее время, затраченное на кластеризацию выделенных

наборов данных в процентах. Алгоритм, которому потребовалось максимальное количество времени Т, имеет 100%, а значения затрат времени других алгоритмов определяются относительно выделенного максимума.

Большое значение в работе иммунных алгоритмов приобретает коэффициент деформации клетки γ , который используется при определении граничных объектов (антител) (5) и при определении уровня мутации объектов (7), (8). Данный коэффициент отображает деформацию клеток организма, которая наблюдается в результате воздействия на них вирусов и бактерий, попавших в данный организм. В предложенных иммунных алгоритмах коэффициент деформации равен 10% ($\gamma = 10\%$). При повышении коэффициента деформации В-клеток предложенные иммунные алгоритмы требуют большего количества проходов цикла кластеризации и уступают стандартному k-means в затратах времени на кластеризацию. Уменьшение коэффициента деформации уменьшает необходимое количество проходов цикла кластеризации, следствием чего является уменьшение затрат времени на кластеризацию, но понижает точность группировки объектов, поэтому его целесообразно устанавливать в диапазоне 5-15%.

Выводы

В работе предложены иммунные модификации алгоритма k-means, в которых используются новые подходы для решения основных задач кластеризации данных. Для сокращения количества вычислений, влияющих на затраты времени, необходимого на кластеризацию, клонированию и мутации подвергаются лишь некоторые объекты исходного множества (граничные антигены, центры В-клеток). Выделение клонируемых объектов производится при использовании значения пороговой аффинности и коэффициента деформации, что значительно уменьшает количество выделяемых объектов и уменьшает затраты времени при кластеризации.

Использование конкурентно-целевого отбора приводит к повышению эффективности отбора клонов при минимальных временных затратах, а определение принадлежности граничных объектов кластерам происходит на основании максимальной средней аффинности его клонов к центральному объекту кластера (центроиду), что значительно ускоряет процесс кластеризации и повышает её точность.

По результатам тестирования алгоритмов на различных наборах данных видно, что предложенные иммунные модификации k-means незначительно усложняют алгоритм кластеризации, однако увеличивают скорость его работы.

Список литературы: 1. *Mirkin, B. G.* Clustering for Data Mining. A Data recovery Approach / B.G. Mirkin. – Taylor & Francis Group, 2005. – 278 p. 2. *Дасгунта, Д.* Искусственные иммунные системы и их применение [Текст]: Пер. с англ. – А.А. Романюха. М.: ФИЗМАТЛИТ, 2006. – 344 с. 3. An Overview of Artificial Immune Systems / J. Timmis, T. Knight, L.N. de Castro, E. Hart. – Natural Computation, 2004. – 29 p. 4. *Duda, R. O.* Pattern classification / R.O. Duda, P.R. Hart, D.G. Stork – Willey & Sons. – 2001. – 738 p. 5. *Кораблёв, Н.М.* Классификация объектов на основе искусственных иммунных систем [Текст] / Н.М. Кораблёв, А.А. Фомичёв // Системы обработки информации. – 2010 – Вып. 6 (87). – С. 13-17.

Поступила в редколлегию 12.07.2011

УДК 004.89

Кластеризация данных методом k-means за помощью штучных иммунных систем / М.М. Кораблёв, О.О. Фомичов // Біоніка інтелекту: наук.-техн. журнал. – 2011. – № 3 (77). – С. 102-106.

У роботі запропоновані імунні підходи до вирішення задачі кластеризації даних методом k-means. Для підвищення швидкодії алгоритму кластеризації використовуються принципи клонування граничних об'єктів та клонування центрів. Проведені експериментальні дослідження запропонованих підходів вказали на високу швидкість кластеризації даних без втрат точності та хорошу масштабованість алгоритмів.

Табл. 2. Бібліогр.: 5 найм.

UDK 004.89

K-means data clustering using artificial immune systems / N.M. Korablev, A.A. Fomichev // Bionics of Intelligence: Sci. Mag. – 2011. – № 3 (77). – P. 102-106.

The paper presents the immune approaches to solving the problem of data clustering method k-means. To speed enhancing operation of clustering algorithm is used principles of boundary objects cloning and center objects cloning of clusters. Experimental studies suggested approaches indicated a high rate of clustering data without loss of accuracy and good scalability algorithms.

Tab. 2. Ref. 5 items.