

## ЭВОЛЮЦИОННАЯ КЛАСТЕРИЗАЦИЯ СЛОЖНЫХ ОБЪЕКТОВ И ПРОЦЕССОВ

В. Снитюк

**Аннотация:** В статье предложен метод кластеризации сложных объектов и процессов, базирующийся на использовании генетического алгоритма. Рассмотрены аспекты его реализации и формирования фитнес-функции. Представлено решение задачи кластеризации областей Украины по социально-экономическим показателям и осуществлен его сравнительный анализ с результатами классических методов.

**Ключевые слова:** Кластеризация, генетический алгоритм.

---

### Введение

Процесс поступательного движения к созданию информационного общества сопровождаются проблемами, связанные с хранением и обработкой больших массивов данных. Их решение связано с интеллектуальным анализом данных, технологии которого формируются на пересечении искусственного интеллекта, статистики, теории баз данных. К ним принадлежат KDD (knowledge discovery in databases) – обнаружение знаний в базах данных, data mining (“раскопка данных”), OLAP (On-line analysis processing) – извлечение информации из многомерных баз данных и другие. Элементы указанных технологий становятся неотъемлемой частью электронных хранилищ данных (Warehouses). Значительную часть информации представляют данные, являющиеся социально-экономическими показателями функционирования сложных систем.

Большим массивам информации свойственно присутствие шумовых эффектов, их обработка приводит к накоплению совокупной ошибки. Для преодоления указанной проблемы необходимо определять значимые факторы и осуществлять их анализ. Уменьшение информационной энтропии может быть также достигнуто путем группировки объектов и извлечения знаний в меньших и функционально связанных совокупностях. Такие процедуры направлены на последовательное преодоление неопределенности. Первым его шагом является решение задачи кластеризации.

---

### Анализ моделей и методов кластеризации

Задача кластеризации заключается в определении групп объектов (процессов), которые являются наиболее близкими один к другому по некоторому критерию. При этом никаких предположений об их структуре, как правило, не делается [Мандель, 1988], [Gorban, 2002]. Большинство методов кластеризации базируется на анализе матрицы коэффициентов сходства, в качестве которых выступают расстояние, сопряженность, корреляция и др. Если критерием или метрикой выступает расстояние, то кластером называют группу точек  $\Omega$ , такую, что средний квадрат внутригруппового расстояния до центра группы меньше среднего расстояния до общего центра в исходном наборе объектов, т.е.  $\bar{d}_{\Omega}^2 < \sigma^2$ , где

$\bar{d}_{\Omega}^2 = \frac{1}{N} \sum_{X_i \in \Omega} (X_i - \bar{X}_{\Omega})^2$ ,  $\bar{X}_{\Omega} = \frac{1}{N} \sum_{X_i \in \Omega} X_i$ . В общем случае, критериями являются:

1. Расстояние Эвклида  $d(X_k, X_l) = \left( \frac{1}{m} \sum_{j=1}^m (X_{kj} - X_{lj})^2 \right)^{\frac{1}{2}}$ .
2. Максимальное расстояние по признакам  $d(X_k, X_l) = \max_{1 \leq j \leq m} |X_{kj} - X_{lj}|$ .
3. Расстояние Махаланобиса  $d(X_k, X_l) = \left[ (X_k - X_l) \cdot R^{-1} \cdot (X_k - X_l)^T \right]^{\frac{1}{2}}$ .
4. Расстояние Хэмминга  $d(X_k, X_l) = \frac{1}{m} \sum_{j=1}^m |X_{kj} - X_{lj}|$ .

Решение задачи минимизации расстояния между объектами равносильно решению задачи минимизации расстояния до объекта, имеющего усредненные характеристики, поскольку, например, для расстояния Хэмминга

$$\sum_{\substack{j=1 \\ k < l}}^m |X_{kj} - X_{lj}| = \sum_{\substack{j=1 \\ k < l}}^m |X_{kj} - \bar{X} + \bar{X} + X_{lj}| \leq \sum_{\substack{j=1 \\ k < l}}^m |X_{kj} - \bar{X}| + \sum_{\substack{j=1 \\ k < l}}^m |X_{lj} - \bar{X}| \leq \sum_{j=1}^m |X_{kj} - \bar{X}| + \sum_{j=1}^m |X_{lj} - \bar{X}| = 2 \sum_{j=1}^m |X_{kj} - \bar{X}|.$$

Задаче кластеризации сопутствуют две проблемы: определение оптимального количества кластеров и получение их центров. Исходными данными для задачи кластеризации являются значения параметров объектов исследования. Очевидно, что определение оптимального количества кластеров является прерогативой исследователя. Предположим, что число кластеров  $K$  задано и  $k \ll m$ , где  $m$  - количество объектов. Получим задачу

$$\sum_{i=1}^K \sum_{j=1}^{m_i} \|X_j - \bar{X}_i\| \rightarrow \min, \quad (1)$$

где  $\bar{X}_i, i = \overline{1, K}$  - среднее значение в кластере,  $\|X_j - \bar{X}_i\|$  - расстояние между объектами. Решением задачи (1) являются центры кластеров  $\bar{X}_i$ , которые могут содержаться среди рассматриваемых объектов, что является достаточно строгим условием, и могут быть представлены любыми точками области исследования.

К традиционным методам кластерного анализа относят древовидную кластеризацию, двухходовое объединение, метод  $K$  средних, метод дендритов, метод корреляционных плеяд и метод шаров [Плюта, 1989]. Преимуществами указанных методов является их простота, инвариантность их техники относительно характера исходных данных и используемых метрик. К недостаткам относят слабую формализованность, что затрудняет применение вычислительной техники, а также низкую точность, следствием чего является предварительные оценки структуры пространства факторов и их информативности. Еще одним методом решения задачи кластеризации является использование самоорганизованной карты Кохонена [Kohonen, 1988]. Проблемой использования такой нейронной сети является выбор начальных весовых коэффициентов, непрерывный характер функционирования и эффективность, оценка которой на сегодняшний день остается проблемой.

В качестве альтернативного метода предлагаем использовать генетический алгоритм.

---

### Генетические алгоритмы – неклассический метод решения задачи оптимизации

---

Первые варианты генетического алгоритма и рассмотрение аспектов его применения появились в работах [Fraser, 1962], [Fraser, 1968], [Bremermann, 1965], [Holland, 1969], [Holland, 1975]. Дальнейшие исследования показали его эффективность в решении инженерных, экономических экологических и других проблем. Главной идеей, лежащей в основе построения генетического алгоритма, является использование идей природного отбора, селекции и мутаций. Его канонический вариант содержит такие этапы:

1. Определение генеральной совокупности особей  $\Theta$ , являющихся потенциальными решениями задачи оптимизации фитнес-функции.
2. Выполнение предварительных шагов алгоритма, заключающихся в определении количества элементов  $K$  выборочной популяции  $\Xi$ , причем  $k \ll |\Theta|$ ; выборе способа нормирования исходных данных; выборе варианта кроссовера, мутации и инверсии, а также соответствующих вероятностей.
3. Для каждого элемента  $\theta_i \in \Xi, i = \overline{1, k}$  вычисляем значения фитнес-функции  $f_i = F(\theta_i)$ .
4. С вероятностями  $P_i^k$ , пропорциональными значением  $f_i$ , выбрать две особи и осуществить кроссовер, вследствие выполнения которого получим две новых особи.
5. С вероятностью  $\frac{1}{2}$  выбираем одну из полученных особей и с вероятностью  $P^m$  осуществляем мутацию.
6. Полученную особь помещаем в новую популяцию  $\Xi^n$ .

7. Повторяем шаги 3-6  $\left\lceil \frac{k}{2} \right\rceil$  раз.

8. Переписываем элементы  $\Xi^n$  в популяцию  $\Xi$ , удаляя старые особи.

Критерием окончания генетического алгоритма могут выступать следующие условия: сходимость элементов популяции  $\Xi$  к одному элементу; максимальное абсолютное отклонение между элементами популяции  $\Xi$  будет меньше некоторого положительного числа  $\delta$ ; максимальное абсолютное отклонения между значениями фитнес-функции будет меньше некоторого малого положительного числа  $\varepsilon$ .

### Формирование фитнес-функции задачи кластеризации

Исходными данными задачи кластеризации являются значения факторов (табл. 1). Предварительно, выполним их нормирование, например, по формуле  $x'_{ij} = \frac{x_{ij} - x_{jmin}}{x_{jmax} - x_{jmin}}$ . Вследствие такого преобразования значения всех факторов будут лежать в единичном гиперкубе  $[0,1]^n$ . Фитнес-функция реализуется следующим алгоритмом:

Шаг 1. Значение фитнес-функции положить равным нулю ( $F = 0$ .)

Шаг 2. Задать количество кластеров  $K$  и указать значение  $m$ .

Шаг 3. Выполнить инициализацию матрицы принадлежности элементов к кластерам  $T_k$ .

Шаг 4. Для всех объектов выполнить следующие шаги. Пусть  $n = 1$

Шаг 5. Вычислить расстояние от  $n$ -го объекта до центров всех  $K$  кластеров, которые является особями из выборочной популяции.

Шаг 6. Среди всех расстояний  $d_j, j = \overline{1, K}$  выбрать минимальное  $d_q$  и отнести  $n$ -й объект к  $q$ -му кластеру. Внести соответствующую запись в матрицу  $T_k$ .

Шаг 7.  $F = F + d_q, n = n + 1$ .

Шаг 8. Если шаги 5-7 выполнены для всех объектов, то получено значение фитнес-функции  $F$ , в противном случае перейти на шаг 5.

Очевидно, что алгоритм получения фитнес-функции можно оптимизировать. Возможность улучшения является его внутренним свойством. Многообразие вариантов операций генетического алгоритма представляют множество внешних свойств процесса получения фитнес-функции. Возможность решения задачи ее оптимизации также предполагает двоичное и десятичное представление исходных данных. И если в первом случае в процедурах генетического алгоритма доминирующим является равномерное распределение, то во втором – при поиске оптимального решения предпочтение отдается значениям, имеющим нормальное распределение с математическим ожиданием, совпадающим с центром кластера. Определение оптимальной дисперсии – еще одна задача, которая остается нерешенной.

Значения факторов исследования				
1	$x_{11}$	$x_{12}$	...	$x_{1n}$
2	$x_{21}$	$x_{22}$	...	$x_{2n}$
...	...	...	...	...
$m$	$x_{m1}$	$x_{m2}$	...	$x_{mn}$

### Кластеризация областей Украины по социально-экономическим признакам

Для проверки эффективности предложенного метода кластеризации были выбраны области Украины. Кластеризация должна была быть осуществлена, исходя из значений социально-экономических показателей. Такими показателями являются:

$X_1$  - валовая прибавочная стоимость в расчете на одного человека (в фактических ценах, грн.);

$X_2$  - территория (тис. кв. км);

$X_3$  - инвестиции в основной капитал на одного человека (в сравнительных ценах, грн.);

$X_4$  - прямые иностранные инвестиции на одного человека (долл. США);

$X_5$  - занятость населения на 10 тыс. человек;

$X_6$  - денежные доходы населения на одного человека (грн.);

$X_7$  - кредиты, предоставленные субъектам хозяйствования на одного человека;

$X_8$  - количество полученных патентов на изобретения на 10 тыс. человек.

В качестве классических методов были выбраны древовидная классификация и метод К средних. Априорно задано два кластера. По методу К средних получены следующие результаты (табл. 2). К первому кластеру отнесены Днепропетровская, Донецкая, Запорожская, Николаевская, Одесская, Полтавская и Харьковская области. Согласно древовидной кластеризации (рис. 1) к первому кластеру отнесены те же области, кроме Донецкой области, хотя она и близка к элементам первого кластера.

Кластеризация была проведена также с использованием эволюционного моделирования. Критерием окончания вычислительного процесса была выбрана максимальное количество итераций равное 1000. Для тех же двух кластеров и восьми факторов количество переменных (хромосома), для которых проводилась оптимизация фитнес-функции, составило 16. В выборочную популяцию вошло двадцать элементов. Учитывая, что фитнес-функция являлась полиэкстремальной, вероятность мутации составила 0, 4. Такое значение увеличило время вычислений, но значительно увеличило точность расчетов за счет выбивания функции из локальных минимумов.

Для контроля за процессом вычислений в режиме реального времени выводилась информация о значении фитнес-функции на каждой итерации (рис.2); о среднем расстоянии между центрами кластеров (рис. 3); значения центров кластеров (рис. 4). Значение фитнес-функции уменьшилось с  $6 \cdot 10^9$  до 11351587, причем на начальных этапах уменьшение происходило гиперболически, а на последних – линейно. Среднее расстояние между центрами кластеров уменьшалось линейно, с постоянно уменьшающейся дисперсией.

В результате вычислений получено два центра кластеров. Координаты первого  $X_1 = 4553, X_2 = 0,01, X_3 = 915, X_4 = 99, X_5 = 4623, X_6 = 2554, X_7 = 791, X_8 = 1,34$ .

Координаты второго  $X_1 = 2952, X_2 = 0,02, X_3 = 530, X_4 = 58, X_5 = 4288, X_6 = 1555,$

$X_7 = 297, X_8 = 0,59$ . К первому кластеру относятся Днепропетровская, Донецкая, Николаевская, Одесская, Полтавская и Харьковская области. Результаты трех рассмотренных методов являются близкими, что свидетельствует о точности эволюционного моделирования. Его преимуществом является также указание центров кластеров и формализация вычислительного процесса. Как было указано выше, предложенная технология может быть усовершенствована.

Таблица 2  
Область Кластер

Область	Кластер
1	2
2	2
3	2
4	1
5	1
6	2
7	2
8	1
9	2
10	2
11	2
12	2
13	2
14	1
15	1
16	1
17	2
18	2
19	2
20	1
21	2
22	2
23	2
24	2

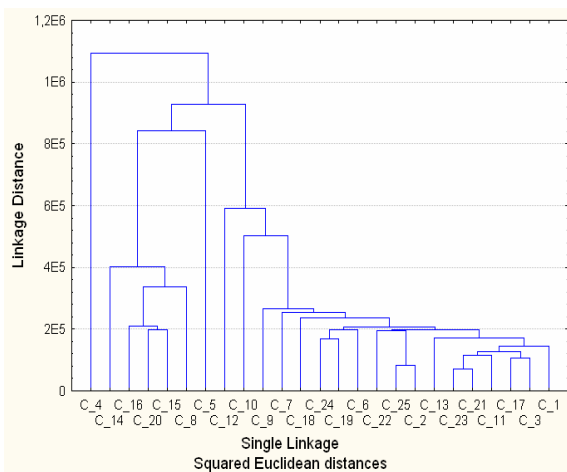


Рис.1 – Результаты древовидной кластеризации

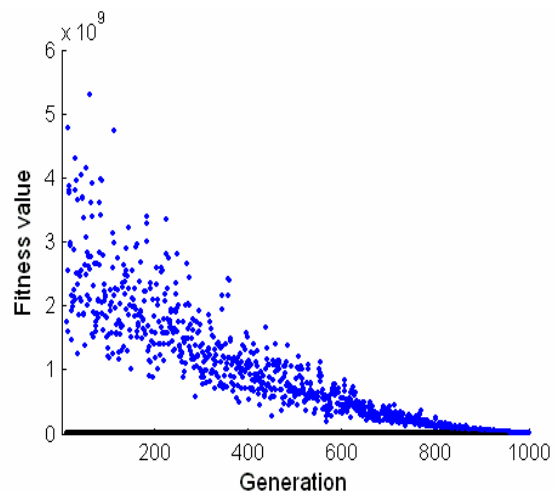


Рис.2 – Значение фитнес-функции

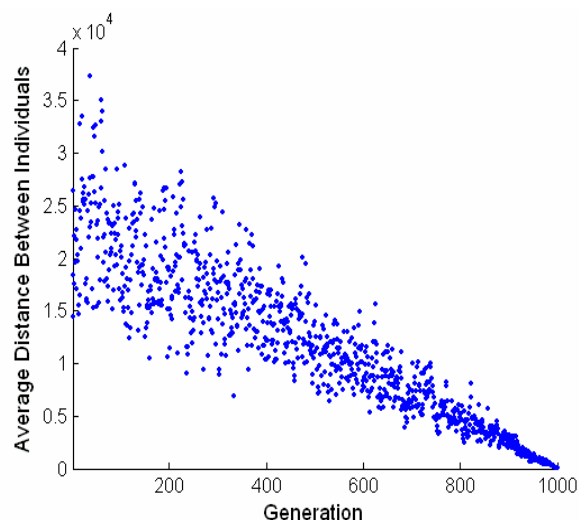


Рис.3 – Расстояние между центрами кластеров

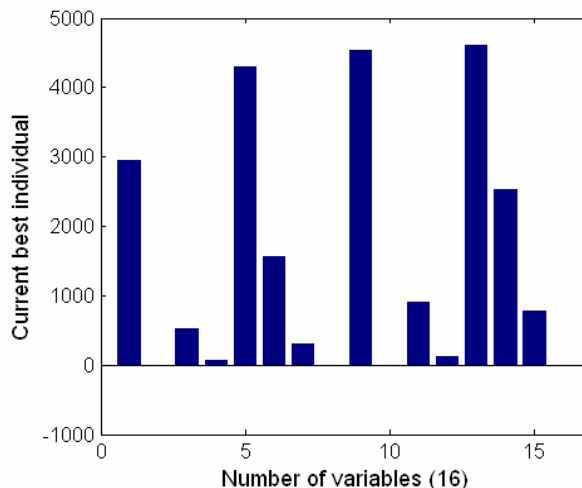


Рис.4 – Координаты центров кластеров

## Заключение

Предложенный метод эволюционного моделирования, базирующийся на использовании генетического алгоритма, эффективно функционирует при обработке массивов большой размерности, поскольку в нем оптимально сочетаются целенаправленный поиск и элементы случайности, направленные на выживание целевой функции из локальных минимумов. Никаких предварительных условий для его использования не требуется. Главным условием оптимизации вычислений является правильная алгоритмизация расчета значений целевой функции. Многовекторность процесса улучшения скорости алгоритма (для генетических алгоритмов особенно актуально) и его точности (поиска глобального минимума фитнес-функции), а также его востребованность свидетельствуют о необходимости решения задачи оптимизации предложенного метода.

## Библиография

- [Мандель, 1988] И.Д. Мандель. Кластерный анализ. Москва, Финансы и статистика, 1988.
- [Gorban, 2002] A.N. Gorban, A.Yu. Zinovyev. Method of Elastic Maps and its Applications in Data Visualization and Data Modeling // Int. Journal of Computing Anticipatory Systems, CHAOS. - 2002. - Vol. 12. - P. 353-369.
- [Плюта, 1989] В. Плюта. Сравнительный многомерный анализ в эконометрическом моделировании. – Москва: Финансы и статистика, 1989.
- [Kohonen, 1988] T. Kohonen. Self-organization and associative memory. – New-York, 2d. ed., Springer Verlag, 1988.
- [Fraser, 1962] A.S. Fraser. Simulation of genetic systems. J. of Theor. Biol., vol. 2, pp. 329-346, 1962.
- [Fraser, 1968] A.S. Fraser. The evolution of purposive behavior. In Purposive Systems, H. von Foerster, J.D. White, L.J. Peterson, and J.K. Russel, Eds. Washington, DC: Spartan Books, pp. 15-23, 1968.
- [Bremermann, 1965] H.J. Bremermann, M. Rogson, S. Salaff. Search by Evolution. In Biophysics and Cybernetic Systems. M. Maxfield, A. Callahan, and L. J. Fogel, Eds. Washington DC: Spartan Books, pp. 157-167, 1965.
- [Holland, 1969] J.H. Holland. Adaptive plans optimal for payoff-only environments. Proc. of the 2nd Hawaii Int. Conf. on System Sciences, pp. 917-920, 1969.
- [Holland, 1975] J.H. Holland. Adaptation in Natural and Artificial Systems. Ann Arbor: Univ. of Michigan Press, 1975.
- [Skurikhin, 1993] A.N. Skurikhin, A.J. Surkan. Identification of parallelism in neural networks by simulation with language J. Proc. of the Intern. Conf. on KPL, APL Quote Quad, Vol.24, No.1, pp.230-237, Toronto, Canada, August 1993.

## Информация об авторе

**Виталий Снитюк** – Киевский национальный университет имени Тараса Шевченко, докторант факультета кибернетики; пр. Акад. Глушкова 2, стр. 6, Киев, Украина; e-mail: [snytyuk@gmail.com](mailto:snytyuk@gmail.com)