# АЛГОРИТМ НЕЧЕТКОЙ КЛАСТЕРИЗАЦИИ В ТЕХНОЛОГИИ DATA MINING

**Гаджиев Ф.Г.[1], Керимов В.А.[2]**

*[1]Гаджиев Фаиг Гасан оглы – кандидат технических наук, доцент;*
*[2]Керимов Вагиф Асад оглы - кандидат технических наук, доцент,*
*кафедра общая и прикладная математика,*
*Азербайджанский государственный университет нефти и промышленности,*
*г. Баку, Азербайджанская Республика*

***Аннотация**: в работе рассматривается проблема разбиения исходного пространства признаков на кластеры, исходя из классических методов кластерного анализа и нечёткой кластеризации, с учётом контекста исходной информации, решение которой формируется с учетом принципов последовательной кластеризации, введения расстояния Хэмминга и оптимизационного подхода.*
***Ключевые слова:** кластерный анализ, целевая функция, функция расстояния, расстояние Хэмминга.*

# FUZZY CLUSTERING ALGORITHM IN TECHNOLOGY DATA MINING

**Hajiyev F.H.[1], Karimov V.A.[2]**

*[1]Hajiyev Faig Hasan oglu - candidate of technical sciences, associate professor;*
*[2]Kerimov Vagif Asad oglu - Candidate of Technical Sciences, Associate Professor,*
*DEPARTMENT OF GENERAL AND APPLIED MATHEMATICS,*
*AZERBAIJAN STATE UNIVERSITY OF OIL AND INDUSTRY,*
*BAKU, REPUBLIC OF AZERBAIJAN*

***Abstract:** the paper considers the problem of partitioning the original feature space into clusters, based on classical methods of cluster analysis and fuzzy clustering, taking into account the context of the original information, the solution of which is formed taking into account the principles of sequential clustering, the introduction of Hamming distance and optimization approach.*
***Keywords:** cluster analysis, target function, distance function, Hamming distance.*

**Введение.** Современное состояние информационных технологий предполагает наличие развитых средств хранения и переработки данных, объемы которых достигают размерности неоднозначной к традиционным представлениям, когда удваивание их числа происходит практически через каждые двух-трёхгодичные интервалы, что значительно усложняет методики их обработки и анализа. Исходя из этого, в последние годы была разработана новая технология Data Mining, предназначенная для решения приведенных задач в процессе системного анализа данных, ориентированных на концептуальные принципы принятия решений на основе не тривиальных, полезных, неизвестных ранее знаний, формализуемых в терминах рассматриваемой проблемной области закономерностями их представления классами, кластерами, ассоциативными правилами, деревьями решений [1].

К настоящему времени широкое использование получила также применение Data Mining для обнаружения и поиска закономерностей в сети Internet , получивший название Web Mining, предполагающий анализ использования Web-ресурсов, извлечение Web-структур и извлечение Web-контента, а ориентация соответствующих ресурсов Web Mining предполагает решение задач описания посетителей сайта, определения типичных сессий и навигационных путей пользователей, а также групп или сегментов посетителей, что связано с задачей кластеризации [2]. При этом вводится понятие Profile Mining, связанное с сегментацией пользователей относительно их идентификации и анализа в аспекте выявления групп пользователей со схожими характеристиками потребностей, желаний и др. [3].

**Постановка задачи.** Предположим, что под {X} мы будем понимать совокупность объектов, {Y}-совокупность идентификаторов кластеров, а $\rho(x, x')$-функцию расстояния между объектами и таким образом, можно говорить о существовании $X^n = \{x_1, x_2, x_3, \dots, x_n\}$, которая должна быть представлена непересекающимися кластерами относительно меры близости или различия соответствующих характеристик с учетом введенной метрики. Классическая задача кластеризации предполагает наличие алгоритма введения функции $d$: X→Y, исходя из которой произвольному объекту x∈X ставится в соответствие идентификатор кластера y∈Y, причем в отдельных ситуациях {Y} может быть заранее известно, а в других оно должно быть определено на основе введенного критерия качества кластеризации [4].

К настоящему времени проводятся исследования по созданию новых принципов кластерного анализа, ориентированных на очень большие базы данных, когда особое значение приобретает вопрос масштабируемости, который, практически, не рассматривался в классических методах.

**Методы решений.** Исследования, проведенные в аспекте указанной постановки задачи показывают ее обусловленность двумя основными факторами: оптимальностью разбиения и определением понятия сходства как краеугольного при решении данного класса задач. Если первый из них рассматривается в аспекте введения функционала, как целевой функции, заданной в виде внутригрупповой суммы квадратов(вск) отклонений:

$$W = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}(\sum_{c=1}^{n} x_i)^2,$$

где $x_i$ – фиксированный признак i-го объекта, то второй, как правило, опирается на функцию расстояния, под которой понимают такие $d(X_i, X_j)$, что будучи неотрицательными и вещественными значениями $d(X_i, X_j)$=0 тогда и только тогда, когда $X_i = X_j$, $d(X_i, X_j)$= $d(X_j, X_i)$ и $d(X_i, X_j) \leq d(X_i, X_к)$+ $d(X_к, X_j)$, $\forall X_i, X_j$ и $X_к \in$El, где под последним понимают l-мерное эвклидово пространство.

При этом могут быть использованы функции расстояния типа: $d_2(X_i, X_j)$ , $d_1(X_i, X_j)$, $d_\infty(X_i, X_j)$ , $d_p(X_i, X_j)$ , $D^2(X_i, X_j)$ [4].

Следует иметь в виду, что при

$$D = \begin{pmatrix} 0 & d_{12} & \dots & d_{1n} \\ d_{21} & d_{22} & \dots & d_{2n} \\ \dots & \dots & \dots & \dots \\ d_{n1} & d_{n2} & \dots & 0 \end{pmatrix}$$

пары значений мер сходства могут определяться, как:

$$S = \begin{pmatrix} 1 & S_{12} & \dots & S_{1n} \\ S_{21} & S_{22} & \dots & S_{2n} \\ \dots & \dots & \dots & \dots \\ S_{n1} & S_{n2} & \dots & 1 \end{pmatrix}$$

где $S_{ij}$ называется коэффициентом сходства, 0≤S($X_i, X_j$)<1, $\forall X_i \neq X_j$, S($X_i, X_i$)=1, S($X_i, X_j$)= S($X_j, X_i$).

Для решения поставленной задачи и с учетом приведенных понятий будем пользоваться методом последовательной кластеризации. Так как A=$\{a_i\}$ $(i = \overline{1,n})$-множество объектов признакового пространства, каждый из них будем рассматривать в виде отдельного элемента: $\{a_1\}$,…,$\{a_n\}$ и среди них определим такие два объекта, которые сходны с учетом введенного критерия сходства и объединим их в один кластер. Тогда новое множество будет состоять из (n-1) кластеров: $\{a_1\}$,…,$\{a_i, a_j\}$, … , $\{a_n\}$, а действуя таким образом получим множество (n-2),(n-3),…,1 с одним кластером.

Предположим, что мерой расстояния может рассматриваться квадрат эвклидовой метрики:

$$D = \begin{array}{c} \\ a_1 \\ a_2 \\ a_3 \\ \dots \\ a_n \end{array} \begin{array}{c} \begin{array}{cccccc} a_1 & a_2 & a_3 & \dots & a_n & \end{array} \\ \begin{bmatrix} 0 & d_{12}^2 & d_{13}^2 & \dots & d_{1n}^2 \\ & 0 & d_{23}^2 & \dots & d_{2n}^2 \\ & & 0 & \dots & d_{3n}^2 \\ & & & \dots & \dots \\ & & & & 0 \end{bmatrix} \end{array}$$

и в то же время объединение $\{a_i, a_j\}$ реализуется, при min расстояния: $d_{i,j}^2$=min{ $d_{i,j}^2$, $i \neq j$}, в результате чего формируется новая матрица расстояний размерностью [(n-1)x(n-1)]:

$$D_1 = \begin{array}{c} \\ \{a_i, a_j\} \\ a_1 \\ a_2 \\ a_3 \\ \dots \\ a_n \end{array} \begin{array}{c} \begin{array}{cccccc} \{a_i, a_j\} & a_1 & a_2 & a_3 & \dots & a_n \end{array} \\ \begin{bmatrix} 0 & d_{ij1}^2 & d_{ij2}^2 & d_{ij3}^2 & \dots & d_{ijn}^2 \\ & 0 & d_{12}^2 & d_{13}^2 & \dots & d_{1n}^2 \\ & & 0 & d_{23}^2 & \dots & d_{2n}^2 \\ & & & 0 & \dots & d_{2n}^2 \\ & & & & \dots & \dots \\ & & & & & 0 \end{bmatrix} \end{array},$$

отображающая процесс представления

$$D, D_1. D_2, \dots, D_r .$$

Поскольку $a_p$ и $a_q$оказываются в одном кластере при $d_{p,q}^2$=2$W_{pq}$ получающий минимальные значения, то $d_{i,q}^2(i = \overline{1,n}, i \neq p)$ в матрице D заменяется на $d_{i,p}^2$=2$W_{ip}$, производится обнуление элементов $q$-ой строки и столбца, $n_p$=$n_p$+$n_q$ , $n_q$=0.

Приведенный подход к кластеризации исходного пространство был реализован с учётом специфики проблемной области на основе следующего алгоритма.

1. Задание контекста исследований.

2. Если решение задачи производится в нечётких условиях, то перейти к

3.Вычисление мер расстояний $d_2, d_1, d_\infty, d_p,\ \boldsymbol{D}^2$ .

4. Перейти к 7.

5. Формирование матрицы, признаков и функций принадлежности, например с использованием экспоненциальных функций лингвистических термов.

6. Введение d $(d_p, d_q)$= $\sum_{j=1}^{n}\left\|\mu\left(x_{pj}\right)-\mu\left(x_{qj}\right)\right\|$, где $x_{pj}-j$-ое значение признака р-го объекта, $\mu\left(x_{pj}\right)-$ j-ое значение функции принадлежности р-го объекта.

7. Выбор и обоснование целевой функции.

8. Формирование требуемых матриц, определение расстояния между кластерами, выбор значения порога.

Следует иметь в виду эффективность нечеткой кластеризации, основанной на оптимизационном подходе, когда с учетом критерия качества Q(P(X)) производится разбиение $P^*$(X)={ $A^1, \dots, A^c$} на «с» нечетких кластеров, с функциями принадлежности $\mu_{li},\ l=\overline{1,c}\ ,\ i=\overline{1,n}$ на множестве объектов X={$x_1, \dots, x_n$}, т.е. Q(P(X))→ $\underset{P(X)\in\Pi}{\text{extr}}$, где под П можно понимать совокупность всех P(X) при ограничениях $\mu_{li}\geq 0,\ \sum_{i=0}^{n}\mu_{li}=1,\ i=\overline{1,n}\ , l=\overline{1,c}$.

**Выводы.** Приведенных в работе подход, основанный на кластеризации с использованием различных подходов был апробирован на материалах, сложности интерпретации которых были проанализированы средствами представленных алгоритмов, ориентированных на стратегию определённой полноты изучения заданного пространства.

### *Список литературы / References*


1. Data Mining – Управление знаниями. sites.google.com/site/upravlenieznaniami
2. BaseGroup Labs.Технология анализа данных.Loginom Company, 2021, basegroup.ru.
3. *Ночевнов Д.* Методы и средства сегментации пользователей Web -сайтов. International Book, Series" Information science and Computing", 2019, pp 99-106.
4. *Дюран Б., Оделл П.* Кластерный анализ. М., Статистика, 1977,123 с.