ИСПОЛЬЗОВАНИЕ В СИСТЕМАХ МОНИТОРИНГА РОБАСТНЫХ АЛГОРИТМОВ КЛАСТЕРИЗАЦИИ

Вишняков А.С.[1], Макаров А.Е.[2], Уткин А.В.[3], Зажогин С.Д.[4], Бобров А.В.[5]

[1]Вишняков Александр Сергеевич – ведущий инженер,
системный интегратор «Крастком»;
[2]Макаров Анатолий Евгеньевич – архитектор решений,
Российская телекоммуникационная компания «Ростелеком»,
г. Москва;
[3]Уткин Александр Владимирович – старший инженер,
Международный системный интегратор «EPAM Systems», г. Минск, Республика Беларусь;
[4]Зажогин Станислав Дмитриевич - старший разработчик,
Международный IT интегратор «Hospitality & Retail Systems»;
[5]Бобров Андрей Владимирович – руководитель группы,
группа технической поддержки,
Компания SharxDC LLC,
г. Москва

Аннотация: *рассмотрены методы построения эффективных алгоритмов кластеризации набора данных в системах мониторинга. С целью создания кластеров с точными границами в условиях наличия выбросов был проведен анализ алгоритмов кластеризации нечетко вероятностного разделения методом нечетких с-средних. Указаны ключевые особенности современного подхода, в рамках которого алгоритмы нечетко-вероятностное разделения набора данных специализируются на обнаружении кластеров имеющих сферическую форму. Показаны преимущества применения нечетко-вероятностного разделения по сравнению с другими моделями кластеризации по методу нечетких с-средних, в том числе, что касается уменьшения требований к вычислительной мощности, необходимой для обработки данных алгоритмов за счет уменьшения количества параметров.*

Ключевые слова: *метод кластеризации нечетких с-средних, нечетко-вероятностное разделение, сферическая форма кластера, робастные процедуры кластеризации.*

УДК 331.225.3

Введение: Разработка надежных и точных алгоритмов кластерного анализа наборов данных имеет широкое применение в современных информационных системах, в частности в системах мониторинга облачных сервисов [1-5]. При этом стабильность или воспроизводимость результатов разделения набора данных в условиях наличия выбросов (outlier data), которые особенно характерны при анализе системами мониторинга потенциально опасных каналов передачи данных, определяется через робастность (выбросоустойчивость) методов кластеризации. На сегодняшний день эффективным подходом автоматической кластеризации считается метод нечетких с-средних Бездека [6-7], однако из-за вероятностных ограничений он является очень чувствительным к выбросам. Поэтому дальнейшие исследования должны быть направлены на нахождение способа нивелирования вероятностных ограничений, что обуславливает ***актуальность исследования*** проведенного в рамках данной работы.

Анализ последних исследований и публикаций в данной области показал приоритет подхода, при котором используются векторы признаков и для потенциальных выбросов создается отдельный класс [8, 9]. Представленный подход получил свое развитие в работах [10-12], где были показаны варианты устойчивых к шуму алгоритмов на базе FCM. С другой стороны, в работах [13, 14] предложен алгоритм вероятностных *c*-средних, который подразумевает разбиение на основе статистических правил. Такой подход является эффективным решением для решения проблемы чувствительности алгоритмов к выбросам, но при этом в результате он часто выявляет совпадающие кластеры [14]. Решение этой проблемы возможно через введение параметра силы отталкивания (repulsive force) между парами кластеров, который прямо пропорционален расстоянию между элементами кластеров [15]. Как было показано данный подход эффективно решает проблему совпадения кластеров, но малоэффективен при работе с наборами данных, для которых кластеры находятся на малом расстоянии друг от друга. Также были предложены алгоритмы нечетко-вероятностных смесей [16, 17], которые показали высокую надежность механизмов кластеризации, но при этом выдавали определенный процент ошибок при наличии экстремальных выбросов [17]. Следует также отметить, что рассмотренные алгоритмы работают с точечными центроидами, которые рассчитываются через усреднение входных данных. Однако при разработке комплексной методологии построения алгоритмов кластеризации необходимо рассмотреть варианты нестандартной формы центроидов кластеров. Так, например, для работы с линейными многообразиями предлагается использовать алгоритм адаптивных нечетких *c*-многообразий [18], аналогично для сферических многообразий можно применять алгоритмы нечеткой *c*-сферической формы (FCSS: fuzzy *c*-spherical shell) [19].

Целью работы, таким образом, стала разработка методологии построения робастных алгоритмов кластеризации, которые эффективно работают с точечными центроидами и центроидами нестандартной формы в условиях экстремальных выбросов.

1. Примеры построения гибридных схем кластеризации на основе метода *c*-средних

Базовым подходом, который используется в алгоритмах кластеризации на основе метода *c*-средних является разделение набора объектов $\{x_n\}$, где $n \in [1;N]$ на C кластеров (каждый -кластер принадлежит к множеству $[1,C]$) в соответствии с принципом минимизации квадратичной целевой функции. В зависимости от типа разделения, который используется при формировании кластеров, можно выделить следующие три группы алгоритмов кластеризации на основе метода *c*-средних:

- методы четкой кластеризации *c*-средних (HCM: hard *c*-means algorithm), в рамках которой используется вероятностное четкое разделение (probabilistic crisp partition);
- методы нечеткой кластеризации *c*-средних (FCM: fuzzy *c*-means algorithm), в рамках которой используется вероятностное нечеткое разделение (probabilistic fuzzy partition);
- метод кластеризации с регуляризацией (PCM: possibilistic *c*-means algorithm), в рамках которой используется вероятностное нечеткое разделение с меньшим количеством ограничений, чем FCM.

Для улучшения отдельных характеристик алгоритмах кластеризации на основе метода *c*-средних при работе с конкретными задачами обычно используют смешанные схемы разделения. В рамках данного исследования предлагается рассмотреть следующие варианты комбинирования подходов FCM и PCM:

- метод кластеризации нечетко-регуляризационных *c*-средних (FPCM: fuzzy-possibilistic *c*-means);
- метод кластеризации регуляризационно-нечетких *c*-средних (PFCM: possibilistic-fuzzy *c*-means).

Целевая функция J для FPCM может быть определена через функцию центроида v_c и функцию вероятностной нечеткой принадлежности (probabilistic fuzzy membership function) $u_{c,n}$:

$$J_{FP} = \sum_{c=1}^{C} \sum_{n=1}^{N} \left(\left(u_{c,n}^{q} + t_{c,n}^{p} \right) \cdot d_{c,n}^{2} \right), \text{т} \qquad (1)$$

где $d_{c,n}, u_{c,n}, d_{c,n}, p$ и q могут быть определены как

$$\begin{cases} d_{c,n} = \|x_n - v_c\| \\ u_{c,n} \in [0;1] \\ d_{c,n} \in [0;1] \\ q > 1 \\ p > 1 \end{cases}. \qquad (2)$$

Функция вероятностной нечеткой принадлежности указывает на уровень принадлежности вектора $\{x_n\}$ кластеру c, где параметры p и q определяют регуляризационную и вероятностную компоненту, соответственно.

Минимизация целевой функции J_{FP} производится через введение следующих ограничений:

$$\sum_{c=1}^{C} u_{c,n} = 1 \text{ для } \forall n; \qquad (3)$$

$$\sum_{n=1}^{N} t_{c,n} = 1 \text{ для } \forall c. \qquad (4)$$

Соответственно, набор функций $u_{c,n}, t_{c,n}$ и v_c может быть определен как:

$$\begin{cases} u_{c,n} = \dfrac{\sqrt[q-1]{\dfrac{1}{d_{c,n}^2}}}{\sum_{f=1}^{C} \sqrt[q-1]{\dfrac{1}{d_{f,n}^2}}} \text{ для } \forall n \text{ и } \forall c \\ t_{c,n} = \dfrac{\sqrt[p-1]{\dfrac{1}{d_{c,n}^2}}}{\sum_{m=1}^{N} \sqrt[p-1]{\dfrac{1}{d_{c,m}^2}}} \text{ для } \forall n \text{ и } \forall c \\ v_c = \dfrac{\sum_{n=1}^{N} \left(\left(u_{c,n}^q + t_{c,n}^p \right) \cdot x_n \right)}{\sum_{n=1}^{N} \left(u_{c,n}^q + t_{c,n}^p \right)} \text{ для } \forall c \end{cases} \quad (5)$$

Основное преимущество FPCM состоит в том, что данном методе не используются штрафные члены, что упрощает процедуру настройки ключевых параметров. Однако, в случае роста входных параметров эффективность регуляризации падает, так при $N \gg C$ FPCM к FCM для любого значения регуляризационной компоненты p.

Аналогичным образом целевая функция PFCM может быть определена через компромиссные параметры (trade-off parameters) a и b,которые используются как коэффициенты регуляризационной и вероятностной компоненты, соответственно, а также штрафного коэффициента η_c:

$$J_{PF} = \sum_{c=1}^{C} \sum_{n=1}^{N} \left(\left(a \cdot u_{c,n}^q + b \cdot t_{c,n}^p \right) \cdot d_{c,n}^2 \right) + \sum_{c=1}^{C} \eta_c \sum_{n=1}^{N} \left(1 - t_{c,n} \right)^p \quad (6)$$

Минимизация целевой функции J_{PF} производится через введение следующих ограничений:

$$\begin{cases} 0 \le u_{c,n} \le 1 \text{ для } \forall n \text{ и } \forall c \\ \sum_{c=1}^{C} u_{c,n} = 1 \text{ для } \forall n \end{cases}; \quad (7)$$

$$\begin{cases} 0 \le t_{c,n} \le 1 \text{ для } \forall n \text{ и } \forall c \\ 0 \le \sum_{c=1}^{C} t_{c,n} \le 1 \text{ для } \forall n \end{cases}. \quad (8)$$

Соответственно уравнения, определяющие минимизацию целевой функции кластеризации и центроиды, могут быть сформулированы как:

$$\begin{cases} u_{c,n} = \dfrac{\sqrt[q-1]{\dfrac{1}{d_{c,n}^2}}}{\sum_{f=1}^{C} \sqrt[q-1]{\dfrac{1}{d_{f,n}^2}}} \text{ для } \forall n \text{ и } \forall c \\ t_{c,n} = \dfrac{1}{1 + \sqrt[p-1]{\dfrac{b \cdot d_{c,n}^2}{\eta_c}}} \text{ для } \forall n \text{ и } \forall c \\ v_c = \dfrac{\sum_{n=1}^{N} \left(\left(a \cdot u_{c,n}^q + b \cdot t_{c,n}^p \right) \cdot x_n \right)}{\sum_{n=1}^{N} \left(b \cdot u_{c,n}^q + b \cdot t_{c,n}^p \right)} \text{ для } \forall c \end{cases} \quad (9)$$

Гибридный алгоритм PFCM можно отнести к робастным и высокоточным, но он остается достаточно чувствительным к экстремальным выбросам.

2. Кластеризация на основе нечетко-вероятностного произведения по методу *c*-средних

Основной проблемой вероятностных алгоритмов кластеризации на основе метода *c*-средних, является то, что появление выброса входного вектора $\{x_n\}$ приводит к высоким значениям функции принадлежности для всех кластеров множества $[1, C]$, что вносит погрешность в расчет центроидов. С другой стороны, для алгоритмов кластеризации с регуляризацией выбросы, напротив, приводят к минимальным значениям функции принадлежности. Обобщенная математическая модель, представленная в предыдущем разделе, позволяет вывести формулу для расчета центроида, которая может быть использована в робастном алгоритме кластеризации [20]:

$$v_c = \frac{\sum_{n=1}^{N} \mu_{c,n}^q \cdot \varphi_{c,n}^p \cdot x_n}{\sum_{n=1}^{N} \mu_{c,n}^q \cdot \lambda_{c,n}^p} \text{ для } \forall c, \quad (10)$$

где параметр $\mu_{c,n}$ описывает вероятностное нечеткое разделение, в то время как параметр $\varphi_{c,n}$ описывает матрицу регуляризованного разделения, которая отвечает за подавление выбросов.

Таким образом, целевая функция кластеризации на основе нечетко-вероятностного произведения по методу *c*-средних (FPPP-FCM: fuzzy-probabilistic product partition fuzzy *c*-means) может быть определена следующим образом:

$$J_{FP3} = \sum_{c=1}^{C} \sum_{n=1}^{N} \left(u_{c,n}^q \cdot \left(t_{n,c}^p \cdot d_{n,c}^2 + \left(1 - t_{n,c}\right)^p \cdot \eta_c \right) \right), \quad (11)$$

при этом используются ограничения приведенные в системах уравнений (7) и (8). Параметры, которые определяют алгоритм FPPP-FCM, включают в себя нечеткую экспоненту $q > 1$, вероятностную экспоненту $p > 1$ и набор условных штрафных коэффициентов $\{\eta_c\}$.

Алгоритм минимизации может быть получен путем использования условий нулевого градиента с помощью функции Лагранжа:

$$L = J_{FP3} + \sum_{n=1}^{N} \lambda_n \cdot \left(1 - \sum_{c=1}^{C} u_{c,n}\right), \qquad (12)$$

где $\{\lambda_n\}$ — набор множителей Лагранжа. Пересечение функцией Лагранжа нулевого уровня, может быть определено через производной функции Лагранжа по $t_{c,n}$ к нулю:

$$u_{c,n}^{q}\left(p \cdot t_{c,n}^{p-1} \cdot d_{c,n}^{2} - \eta_c \cdot p\left(1 - t_{c,n}\right)^{p-1}\right) = 0. \qquad (13)$$

Как можно видеть при $u_{c,n} = 0$ значение $t_{c,n}$ может быть любым. Соответственно при $u_{c,n} \neq 0$:

$$\frac{1}{t_{c,n}} - 1 = \sqrt[p-1]{\frac{d_{c,n}^{2}}{\eta_c}} \rightarrow t_{c,n} = \frac{1}{1 + \sqrt[p-1]{\frac{d_{c,n}^{2}}{\eta_c}}} \text{ для } \forall n \text{ и } \forall c\,. \qquad (14)$$

Аналогично, приравнивание производной функции Лагранжа по $u_{c,n}$ к нулю дает следующее уравнение:

$$q \cdot u_{c,n}^{q-1}\left(t_{c,n}^{p} \cdot d_{c,n}^{2} - \eta_c \cdot \left(1 - t_{c,n}\right)^{p}\right) = \lambda_n \rightarrow$$

$$\rightarrow u_{c,n}^{q} = \sqrt[p-1]{\frac{\lambda_n}{q \cdot \left(t_{c,n}^{p-1} \cdot d_{c,n}^{2} - \eta_c \cdot \left(1 - t_{c,n}\right)^{p}\right)}}\,. \qquad (15)$$

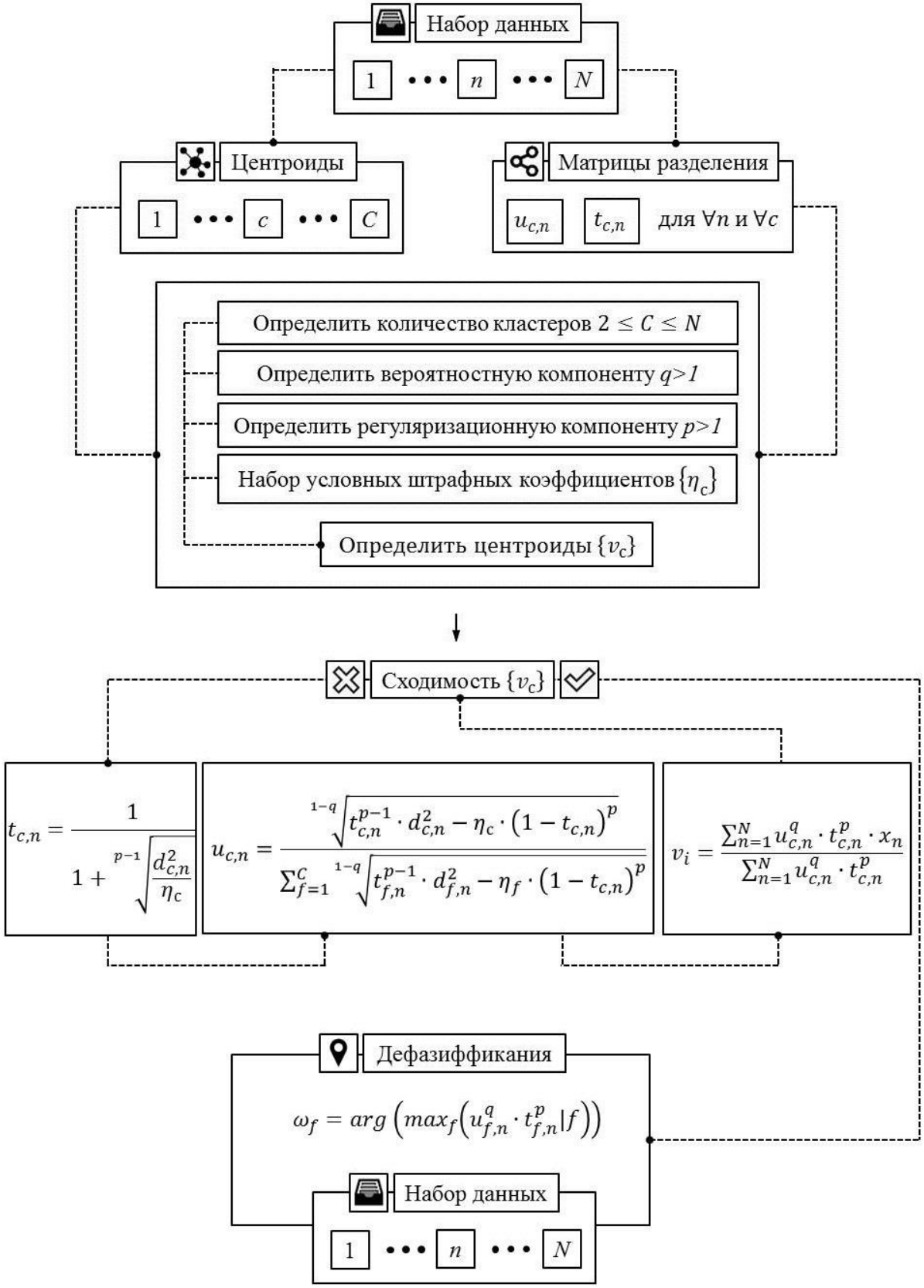

Рис. 1. Алгоритм кластеризации на основе нечетко-вероятностного произведения по методу с-средних

Соответственно:

$$u_{c,n} = \frac{\sqrt[1-q]{t_{c,n}^{p-1} \cdot d_{c,n}^2 - \eta_c \cdot \left(1 - t_{c,n}\right)^p}}{\sum_{d=1}^{C} \sqrt[1-q]{t_{d,n}^{p-1} \cdot d_{d,n}^2 - \eta_d \cdot \left(1 - t_{c,n}\right)^p}} \text{ для } \forall n \text{ и } \forall c. \qquad (16)$$

И, наконец, приравнивание производной функции Лагранжа по функции центроида к нулю дает следующее уравнение:

$$-2\sum_{n=1}^{N}\left(u_{c,n}^q \cdot t_{c,n}^p (x_n - v_c)\right) = 0 \;\rightarrow\; v_i$$

$$= \frac{\sum_{n=1}^{N} u_{c,n}^q \cdot t_{c,n}^p \cdot x_n}{\sum_{n=1}^{N} u_{c,n}^q \cdot t_{c,n}^p} \text{ для } \forall c. \qquad (17)$$

Финальный процесс дефаззификации состоит в присвоении каждого x_n кластеру c с индексом ω_f:

$$\omega_f = arg\left(max_f\left(u_{f,n}^q \cdot t_{f,n}^p | f\right)\right), \qquad (18)$$

В случае равных элементов множества $\{\eta_c\}$ данное уравнение может быть существенно упрощено:

$$\omega_f = arg\left(max_f\left(d_{f,n} | f\right)\right), \qquad (19)$$

где $f \in [1; C]$.

3. Формирование сферических центроидов при кластеризация на основе нечетко-вероятностного произведения по методу *c*-средних

При формировании сферических центроидов (сфероидов) при кластеризация на основе нечетко-вероятностного произведения по методу *c*-средних используется прежнее определение функции кластеризации J_{FP3} с дополнительным уточнением значения $d_{i,k}$ [21]:

$$\begin{cases} J_{FP3} = \sum_{c=1}^{C}\sum_{n=1}^{N}\left(u_{c,n}^q \cdot \left(t_{n,c}^p \cdot d_{n,c}^2 + \left(1 - t_{n,c}\right)^p \cdot \eta_c\right)\right), \\ d_{n,c} = \left|\|x_n - \theta_c\|^2 - r_c^2\right| \end{cases} \qquad (20)$$

где θ_c — центр сфероида, а r_c — радиус сфероида.

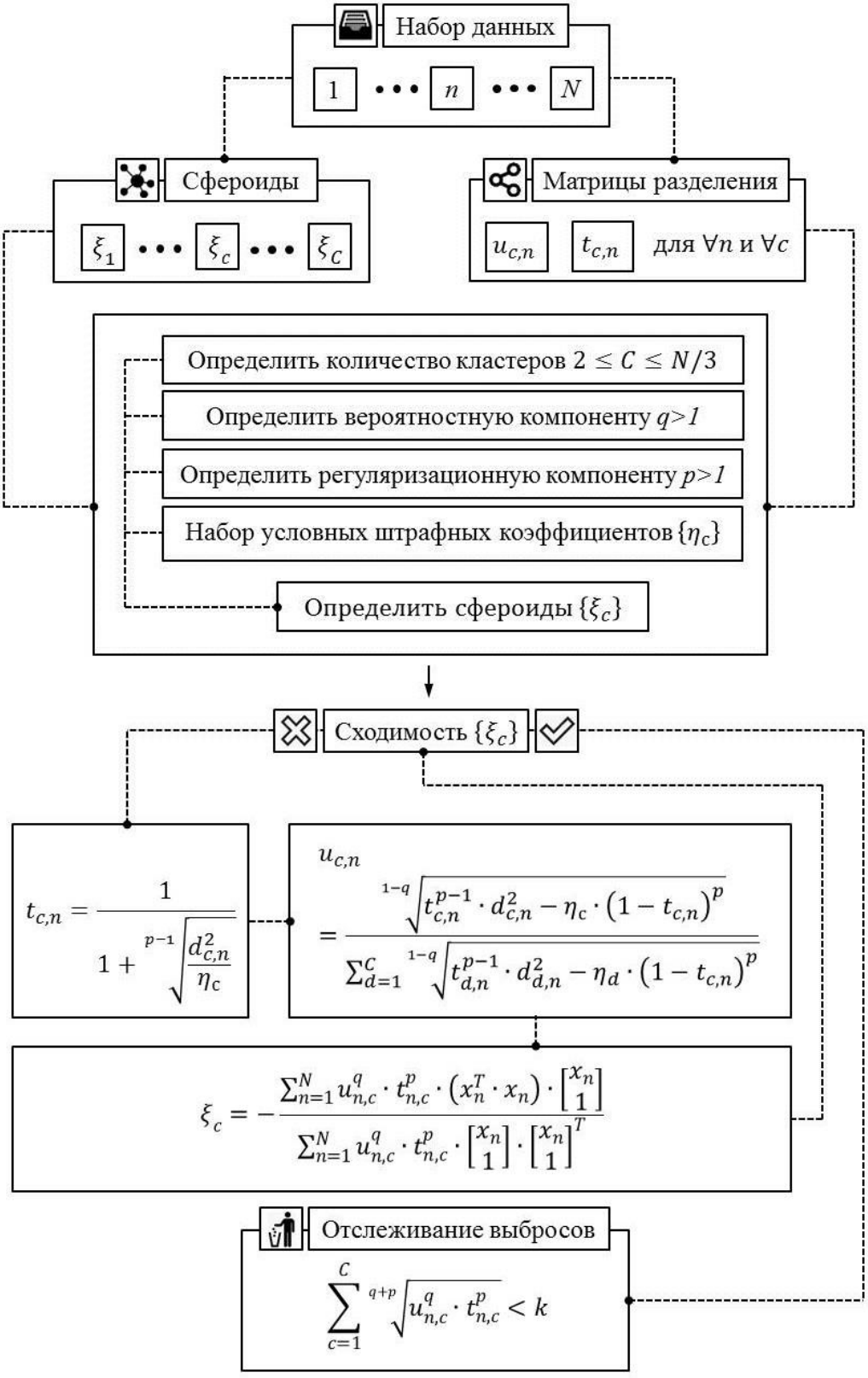

Рис. 2. Алгоритм формирования сфероидов при кластеризации на основе нечетко-вероятностного произведения по методу c-средних

На рис. 2 показан алгоритм формирования сфероидов при кластеризации на основе нечетко вероятностного произведения по методу *c*-средних. Для получения алгоритма минимизации целевой функции в данном случае также могут быть использованы уравнения для нулевого градиента множителя

Лагранжа. Таким образом, для описания сфероидов применяется математический аппарат, описывающий кластеризацию, приведенный в формулах (13) и (16) с уточнением значения $d_{n,c}$.

Однако для оптимизации алгоритма поиска сфероидов необходимо оптимизировать и определение $d_{n,c}$ через функцию сфероида ξ_c:

$$\begin{cases} d_{n,c} = \sqrt{\xi_c^T \cdot \begin{bmatrix} x_n \\ 1 \end{bmatrix} \cdot \begin{bmatrix} x_n \\ 1 \end{bmatrix}^T \cdot \xi_c + 2x_n^T \cdot x_n \cdot \begin{bmatrix} x_n \\ 1 \end{bmatrix}^T \cdot \xi_c + (x_k^T \cdot x_k)^2} \\ \xi_c = \begin{bmatrix} -2\theta_c \\ \theta_c^T \cdot \theta_c - r_c^2 \end{bmatrix} \end{cases} \quad (21)$$

Приравнивание производной функции Лагранжа по функции сфероида к нулю дает следующее уравнение:

$$2\sum_{n=1}^{N} u_{c,n}^m \cdot t_{c,n}^p \cdot \left(\begin{bmatrix} x_n \\ 1 \end{bmatrix} \cdot \begin{bmatrix} x_n \\ 1 \end{bmatrix}^T \cdot \xi_c + (x_n^T \cdot x_n) \cdot \begin{bmatrix} x_n \\ 1 \end{bmatrix} \right) = 0. \quad (22)$$

Что позволяет определить функцию сфероида ξ_c:

$$\xi_c = -\frac{\sum_{n=1}^{N} u_{n,c}^q \cdot t_{n,c}^p \cdot (x_n^T \cdot x_n) \cdot \begin{bmatrix} x_n \\ 1 \end{bmatrix}}{\sum_{n=1}^{N} u_{n,c}^q \cdot t_{n,c}^p \cdot \begin{bmatrix} x_n \\ 1 \end{bmatrix} \cdot \begin{bmatrix} x_n \\ 1 \end{bmatrix}^T}. \quad (23)$$

Факт наличия выбросов при этом отслеживается через соотнесение $u_{c,n}$ и $t_{c,n}$ коэффициентом:

$$\sum_{c=1}^{C} \sqrt[q+p]{u_{n,c}^q \cdot t_{n,c}^p} < k, \quad (24)$$

где k выбирается арбитрарно.

Таким образом, разработанная комплексная методология дает эффективный инструментарий для построения робастных алгоритмов кластеризации, которые работают как с точечными центроидами, так и с центроидами нестандартной формы (в том числе в условиях экстремальных выбросов).

Выводы

В результате проведенного анализа были предложены гибридные алгоритмы кластеризации на основе нечетко вероятностного произведения по методу *c*-средних в частности:

1. алгоритм кластеризации на основе нечетко вероятностного произведения по методу *c*-средних;
2. алгоритм формирования сфероидов при кластеризации на основе нечетко вероятностного произведения по методу *c*-средних.

Предложенная методология включает в себя математический аппарат для построения высокоточных, робастных алгоритмов кластеризации наборов данных, которые могут быть использованы при мониторинге информационных систем.

Список литературы

1. *Lee S., Kim J. & Jeong Y.*, 2017. Various Validity Indices for Fuzzy K-means Clustering. Korean Management Review, 46(4), 1201-1226.

2. *Chen S.*, 2017. An improved fuzzy decision analysis framework with fuzzy Mahalanobis distances for individual investment effect appraisal. Management Decision, 55 (5), 935-956.
3. *Lewis R.H., Paláncz B. & Awange J.*, 2015. Application of Dixon resultant to maximization of the likelihood function of Gaussian mixture distribution. ACM Communications in Computer Algebra, 49(2), 57-57.
4. *Kumar P. & Chaturvedi A.*, 2016. Probabilistic query generation and fuzzy c-means clustering for energy-efficient operation in wireless sensor networks. International Journal of Communication Systems. 29 (8), 1439-1450.
5. *Raveendran R. & Huang B.*, 2016. Mixture Probabilistic PCA for Process Monitoring - Collapsed Variational Bayesian Approach. IFAC-PapersOnLine, 49 (7). 1032-1037.
6. *Hathaway R.J., Overstreet D.D., Murphy T.E. & Bezdek J.C.*, 2001. Relational data clustering with incomplete data. Applications and Science of Computational Intelligence IV.
7. *Hathaway R., Huband J. & Bezdek J.* (n.d.). Kernelized Non-Euclidean Relational Fuzzy c-Means Algorithm. The 14th IEEE International Conference on Fuzzy Systems, 2005. FUZZ 05.
8. *Davé R.N.* Characterization and detection of noise in clustering. Patt. Recogn. Lett. 12, 657–664 (1991).
9. *Klawonn F.*, 2004. Noise Clustering with a Fixed Fraction of Noise. Applications and Science in Soft Computing. 133-138.
10. *Menard M., Damko C., Loonis P*. The fuzzyc+2 means: solving the ambiguity rejection in clustering. Patt. Recogn. 33, 1219–1237, 2000.
11. *Xu H. & Yue X.*, 2009. An Adaptive Fuzzy Switching Filter for Images Corrupted by Impulse Noise, 2009. Sixth International Conference on Fuzzy Systems and Knowledge Discovery.
12. *Alanzado A.C., Miyamoto S.* Fuzzyc-means clustering in the presence of noise cluster for time series analysis. Proc. Modeling Decisions in Artificial Intelligence (MDAI), Lect. Notes Comp. Sci. 3558, 156–163 (2005)
13. *Nasraoui O. & Krishnapuram R.* (n.d.). A novel approach to unsupervised robust clustering using genetic niching. Ninth IEEE International Conference on Fuzzy Systems. FUZZ- IEEE 2000 (Cat. No.00CH37063).
14. *Hamasuna Y., Endo Y. & Miyamoto S.*, 2009. On tolerant fuzzy c-means clustering and tolerant possibilistic clustering. Soft Computing, 14 (5), 487-494.
15. *Timm H., Borgelt C., Döring C., Kruse R.* An extension to possibilistic fuzzy cluster analysis. Fuzzy Sets and Systems 147, 3–16, 2004.
16. *Pal N.R., Pal K., Keller J.M., Bezdek J.C.* A possibilistic fuzzy c-means clustering algorithm. IEEE Trans. Fuzzy Syst. 13, 517–530, 2005.
17. *Szilágyi L.* Fuzzy-Possibilistic Product Partition: a novel robust approach to c-means clustering. Proc. Modeling Decisions in Artificial Intelligence (MDAI), Lect. Notes Comp. Sci. 6820, 150–161, 2011.
18. *Suhaili S.M., Jambli M.N. & Huspi S.H.*, 2011. Evaluation of FCV and FCM clustering algorithms in cluster-based compound selection, 2011. 7th International Conference on Information Technology in Asia.
19. *Wang T. & Shen Q.*, 2002. Fuzzy C spherical shells cluster algorithm and an application to blood cell image. Second International Conference on Image and Graphics.
20. *Szilágyi L., Szilágyi S.M., Benyó B., Benyó Z.* Intensity inhomogeneity compensation and segmentation of MR brain images using hybridc-means clustering models. Biomed. Sign. Proc. Contr. 6, 3–12. 2011.
21. *Szilágyi L.* Robust spherical shell clustering using fuzzy-possibilistic product partition. Int. J. Intell. Syst. 28, 524–539, 2013.