ПОСТРОЕНИЕ НЕЧЕТКИХ АЛГОРИТМОВ ПОЛУАВТОМАТИЧЕСКОГО ОБУЧЕНИЯ НА ОСНОВЕ МАТРИЦ РАЗЛИЧИЙ И ЯДЕРНЫХ МАТРИЦ

Усов А.Е.[1], Варламов А.А.[2], Бабкин О.В.[3], Дос Е.В.[4], Мостовщиков Д.Н.[5]

[1]Усов Алексей Евгеньевич – ведущий архитектор;
[2]Варламов Александр Александрович – старший архитектор;
[3]Бабкин Олег Вячеславович – старший архитектор;
[4]Дос Евгений Владимирович – архитектор;
[5]Мостовщиков Дмитрий Николаевич – старший архитектор,
системный интегратор «Li9 Technology Solutions»,
г. Райли, Соединенные Штаты Америки

Аннотация: *рассмотрены методы применения полуавтоматической кластеризации в практической задаче обработки наборов частично помеченных данных. Проведен анализ алгоритмов, использующих жесткие ограничения по наличию и отсутствию определенных типов данных в кластере. Показан приоритет современного подхода, в рамках которого предлагается использовать полуавтоматическую кластеризацию с мягкими попарными ограничениями. В основу данного подхода было предложено положить алгоритмы, которые базируются на методе нечетких c-средних. В частности, для решения поставленной задачи с точки зрения мягких ограничений были модифицированы алгоритмы энтропийной регуляризованной кластеризации c-средних и неопределенной ядерной кластеризации c-средних. Также был предложен подход, который включает в алгоритм попарные ограничения в том случае, когда мягкие ограничения не дают достаточного уровня эффективности кластеризации набора данных.*

Ключевые слова: *полуавтоматическая кластеризация, метод нечетких c-средних, метод энтропийной кластеризации c-средних, метод неопределенной ядерной кластеризации c-средних, bFCM, eFCM, RFCM.*

УДК 331.225.3

Введение: Автоматический кластерный анализ больших наборов данных через построения групп объектов основании параметров, определяющих их сходство, активно используется в современных информационных системах [1-10]. Следует отметить, что при решении современных практических задач обработки наборов частично помеченных данных более эффективно использовать методы полуавтоматической кластеризации наборов частично помеченных данных, что обуславливает **актуальность исследования** проведенного в рамках данной работы.

Анализ последних исследований и публикаций в данной области показал приоритет метода нечетких *c*-средних (FCM: Fuzzy *c*-means) и алгоритмов на его основе [6-10], в первую очередь метода нечетких *c*-средних Бездека (bFCM: Bezdek type FCM). Кроме того были рассмотрены алгоритмы FCM, которые основываются на этропийной регуляризации (eFCM: entropy-regularized FCM) и, соответственно, могут на математическом уровне комбинироваться с bFCM [11-12]. Другим вариантом развития bFCM является реляционная кластеризация нечетких *c*-средних Бездека (bRFCM: Bezdek-type relational fuzzyc-means) [13], в рамках которой реляционная модель используется для количественного определения связей между парами объектов. Данная парадигма была расширена для неевклидовой реляционной модели (NEbRFCM: non-Euclidean bRFCM), которая работает с соответствующими типами данных через расчет различий между ними [14]. Комбинирование eFCM и bRFCM [15-18] позволил построить метод энтропийной регуляризованной кластеризации *c*-средних (eRFCM: entropy-regularized relational fuzzy *c*-means), энтропийной регуляризованной ядерной кластеризации *c*-средних (K-bFCM: entropy-regularized kernel fuzzy *c*-means) и ядерной кластеризации *c*-средних Бездека (K-eFCM: entropy-regularized kernel fuzzy *c*-means) .

Далее были рассмотрены методы полуавтоматической кластеризации *c*-средних, которые показывают свою эффективность при работе с данными, часть из которых не имеет меток [19-27]. Был проведен анализ использования мягких и жестких ограничений для данной концепции кластеризации, в частности использование попарных ограничений [23-25] .

Целью работы, таким образом, стала разработка методологии построения комплексных алгоритмов на основе метода кластеризации *c*-средних, которая комбинирует подходы, применяемые в алгоритмах bRFCM, eRFCM, NEbRFCM, K-bFCM, K-eFCM и IK-bFCM в соответствии с типом поставленной задачи.

1. Основы построения алгоритмов полуавтоматической кластериза-ции на основе метода нечетких *c*-средних

Для построения методологии, которая объединяет принципы алгоритмов bRFCM, eRFCM, NEbRFCM, K-bFCM, K-eFCM и IK-bFCM следует определить ключевые элементы данных методов, указывая на общие подходы и отличия. В частности при построении базовой модели можно выделить следующие компоненты (рис. 1):

- набор данных, который рассматривается как множество элементов $\{x_n\}$, где $n \in [1;N]$;
- матрица различий D между элементами набора данных (dissimilarity data matrix) для bRFCM, eRFCM и NEbRFCM, которая определяется через множество вещественных чисел $\mathbb{R}$ и N, где $D \in \mathbb{R}^{N \times N}$;
- ядерная матрица K (kernel matrix) для bRFCM, eRFCM и NEbRFCM, которая определяется через множество вещественных чисел $\mathbb{R}$ и N, где $K \in \mathbb{R}^{N \times N}$;
- блочная матрица $u_{i,k}$, которая определяет разделение набора данных на i кластеров, где $i \in [1,I]$, где $u_{i,n} \in \mathbb{R}^{I \times N}$.

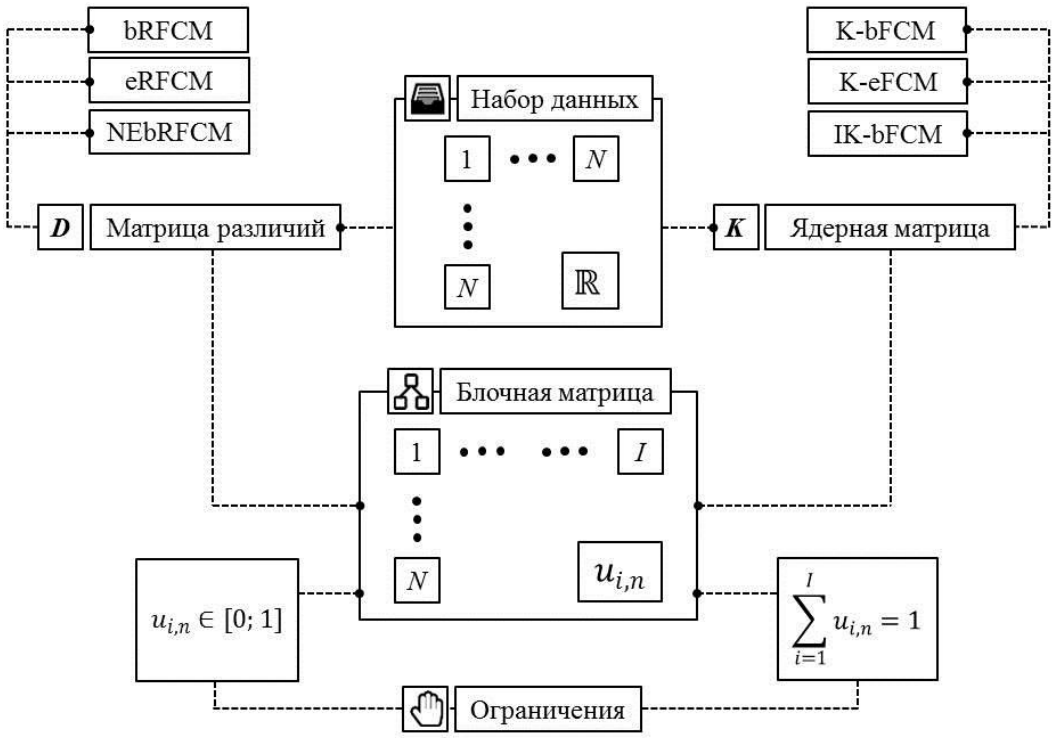

Рис. 1. Базовый алгоритм полуавтоматической кластеризации на основе методов нечетких c-средних Бездека

Рассмотрим ограничения для $u_{i,k}$, которые будут актуальны для всех перечисленных алгоритмов (bRFCM, eRFCM, NEbRFCM, K-bFCM, K-eFCM и IK-bFCM):

$$\begin{cases} u_{i,n} \in [0;1] \\ \sum_{i=1}^{I} u_{i,n} = 1 \end{cases} \quad (1)$$

весовой показатель *m* определяет уровень нечеткости алгоритма. Таким образом, при $m \to 1$ модель приближается к четкому *c*-разделению, а при $m \to \infty$ значение $u_{i,n}^{m} \to 1/C$ для любых объектов и кластеров объектов.

Первый предложенный алгоритм комбинирует методы bRFCM и eRFCM, поэтому математический аппарат в данном случае включает коэффициенты λ и m, которые определяются следующим образом:

- λ — коэффициент ограничения фазификации (fuzzification penalty), где под фаззификацией подразумевается подготовка задачи для решения методами нечеткой логики;
-
- m — весовой показатель, который также определяет уровень фаззификации алгоритма.

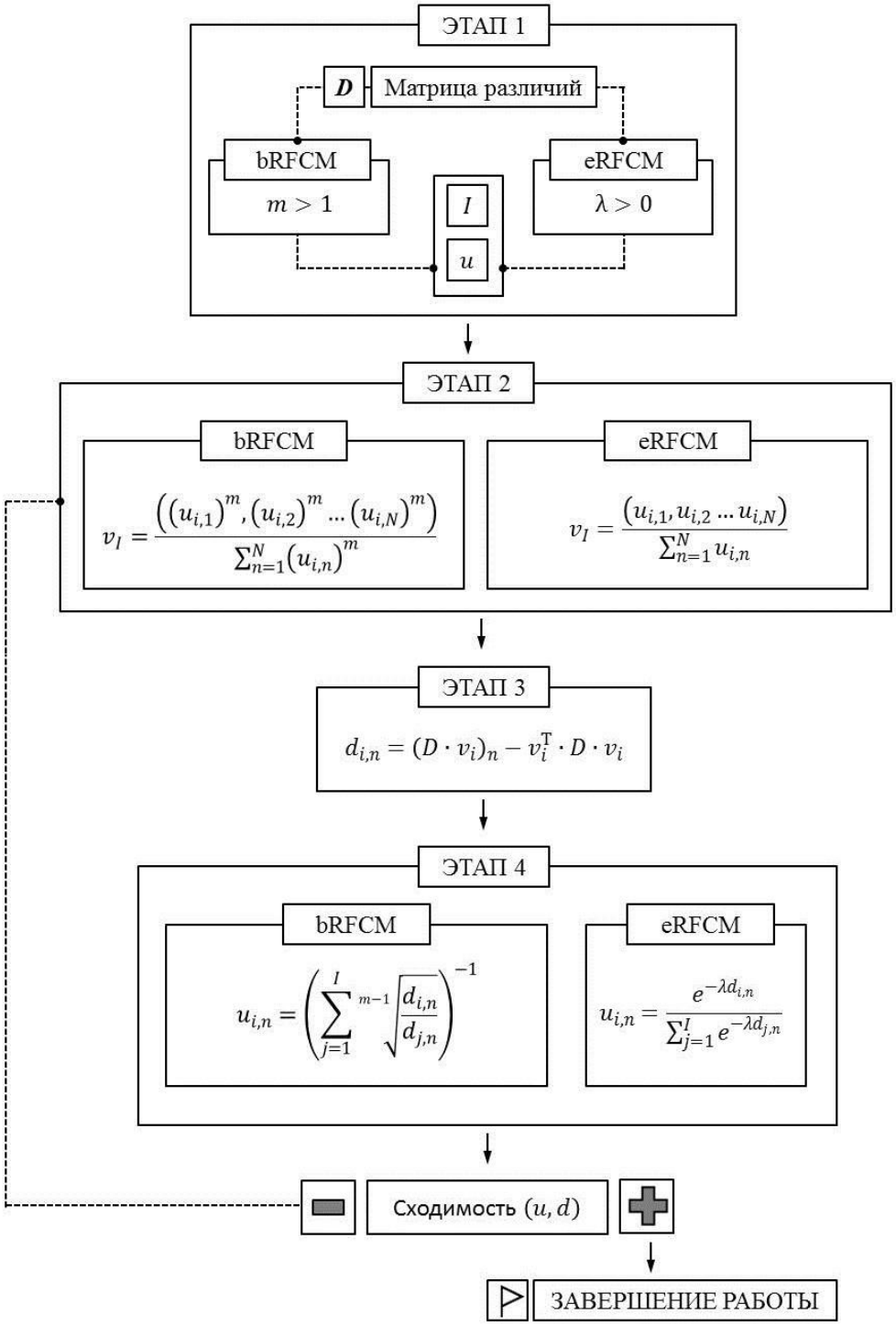

Рис. 2. Алгоритм кластеризации на основе методов bRFCM и eRFCM

При росте λ значение $u_{i,n} \to 1/I$ для всех объектов и кластеров объектов. Аналогично, при $m \to 1$ модель приближается к четкому *c*-разделению, а при $m \to \infty$ значение $u_{i,n} \to 1/I$ для всех объектов и кластеров объектов. Алгоритм, который комбинирует методы bRFCM и eRFCM, включает в себя пять этапов (рис. 2):

1. За основу берется матрица различий D, причем для bRFCM устанавливается значение $m > 1$, а для eRFCM — $\lambda > 0$. Далее определяется значение I и устанавливается функция принадлежности u.
2. Рассчитывается $v_i(m, u_{i,n})$ для bRFCM и $v_i(\lambda, u_{i,n})$ для eRFCM.
3. Рассчитывается $d_i\left(v_i(m, u_{i,n})\right)$ для bRFCM и $d_i\left(v_i(\lambda, u_{i,n})\right)$ для eRFCM.
4. Рассчитывается $u_{i,n}\left(d_i\left(v_i(m, u_{i,n})\right)\right)$ для bRFCM и $u_{i,n}\left(d_i\left(v_i(\lambda, u_{i,n})\right)\right)$ для eRFCM.
5. Если определена сходимость (u, d), то алгоритм завершается. В противном случае — переход к этапу «2».

Представленный алгоритм является простым комбинированием алгоритмов bRFCM и eRFCM, но при этом он может быть положен в основу широкого класса комплексных алгоритмов кластеризации по методу нечетких *c*-средних.

3. Алгоритмы полуавтоматической кластеризации на основе метода нечетких *c*-средних Бездека

Предложенный выше алгоритм работает с евклидовой метрикой, т.е. элементы матрицы различий для множества объектов $\{x_1, x_2 \dots x_N\}$ рассчитываются как $D_{n,n'} = \|x_n - x_{n'}\|_2^2$. Алгоритм для неевклидовой метрики может совпадать с предыдущим алгоритмом на уровне первых двух этапов, но на третьем этапе он даст отрицательное значение $d_{i,n}$ для определенных значений I и m. Таким образом, для неевклидовой метрики не всегда выполняется условие $u_{i,n} \in [0; 1]$, указанное в уравнении (1). Поэтому в данном случае было предложено комбинировать алгоритмы NEbRFCM и eRFCM (рис. 3):

1. Для матрицы различий D, определяются значения $m > 1$(алгоритм bRFCM) и $\lambda > 0$ (алгоритм eRFCM). Определяется I и u, а бета-распределение устанавливается как $\beta = 0$.
2. Рассчитывается $v_i(m, u_{i,n})$ для bRFCM и $v_i(\lambda, u_{i,n})$ для eRFCM.
3. Рассчитывается $d_i\left(v_i(m, u_{i,n})\right)$ для bRFCM и $d_i\left(v_i(\lambda, u_{i,n})\right)$ для eRFCM.
4. В том случае, если $d_{i,n} < 0$, рассчитывается значение $\Delta\beta$ и в соответствии с ним пересчитывается $d_{i,n}$ и β.
5. Рассчитывается $u_{i,n}$.
6. Если определена сходимость (u, d), то алгоритм завершается. В противном случае — переход к этапу «2».

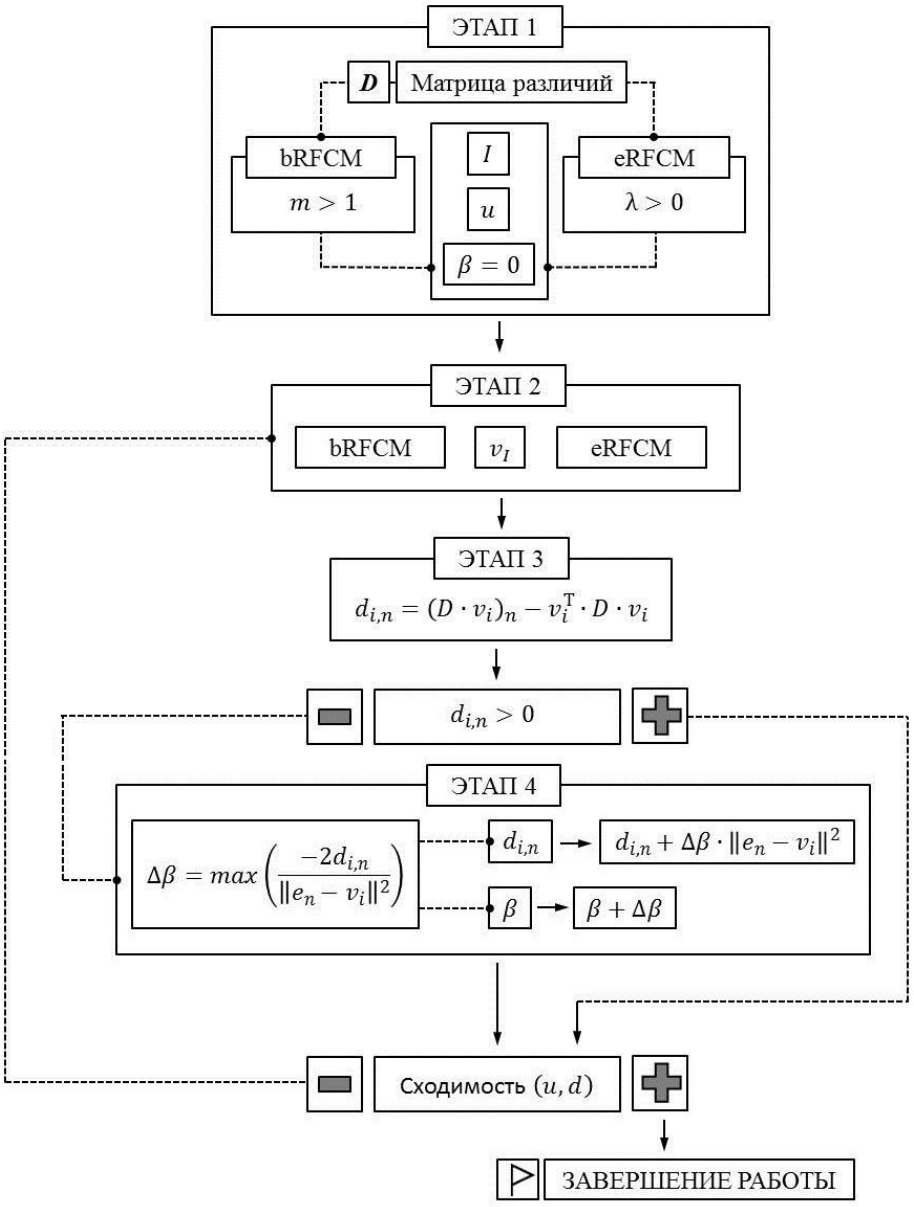

Рис. 3. Алгоритм кластеризации на основе методов bRFCM и eRFCM для неевклидовой метрики

Аналогично может быть построен алгоритм, совмещающий методы K-bFCM и K-eFCM. Центры кластеризации W_i^b (для K-bFCM) и W_i^b (для K-eFCM) при этом рассчитываются как:

$$\begin{cases} W_i^b = \dfrac{\left((u_{i,1})^m,(u_{i,2})^m \ldots (u_{i,N})^m\right)^{\mathrm{T}}}{\sum_{n=1}^{N}(u_{i,n})^m} \\ W_i^e = \dfrac{(u_{i,1},u_{i,2} \ldots u_{i,N})^{\mathrm{T}}}{\sum_{n=1}^{N} u_{i,n}} \end{cases}. \qquad (2)$$

Соответственно, алгоритм, совмещающий методы K-bFCM и K-eFCM, включает в себя следующие этапы (рис. 4):

1. Определяется количество кластеров I весовой показатель фазиффикации $m > 1$ для K-bFCM и $\lambda > 0$ для K-eFCM.

2. Обновляются центры кластеризации в соответствии с уравнением (2).

3. Рассчитывается степень различия $d_{i,n}$ между элементами набора данных и центрами кластеризации.

4. Обновляется функция принадлежности $u_{i,n}$ для K-bFCM и для K-eFCM.

5. Если определена сходимость (u, d, W), то алгоритм завершается. В противном случае — переход к этапу «2».

Метод K-bFCM основывается на том, что ядерная матрица K является положительно полуопределённой. Таким образом, метод K-bFCM может работать с неопределенной K, если количество отрицательных собственных значений минимально, что вносит в метод специфическое ограничение. Чтобы преодолеть это ограничение, было предложено использовать бета-распределение при преобразовании ядерной матрицы [18]:

$$K' = K + \beta \cdot E \qquad (3)$$

где K' будет определена как положительная полуопределённая для существенно большего количества наборов, если $\beta > 0$.

Соответствующий алгоритм (рис. 5) этап предварительного определения количества кластеров, весового показателя и ядерной матрицы, обновление центов кластеризации в соответствии с уравнением (2), расчет степени различия между элементами набора данных и центрами кластеризации и его пересчет в случае, если $d_{i,n}$<0. Далее обновляется функция принадлежности (отдельно для для K-bFCM и для K-eFCM) и если определена сходимость (u, d, W), то алгоритм завершается и выдает полученный результат, а в противном случае осуществляется переход к этапу «2».

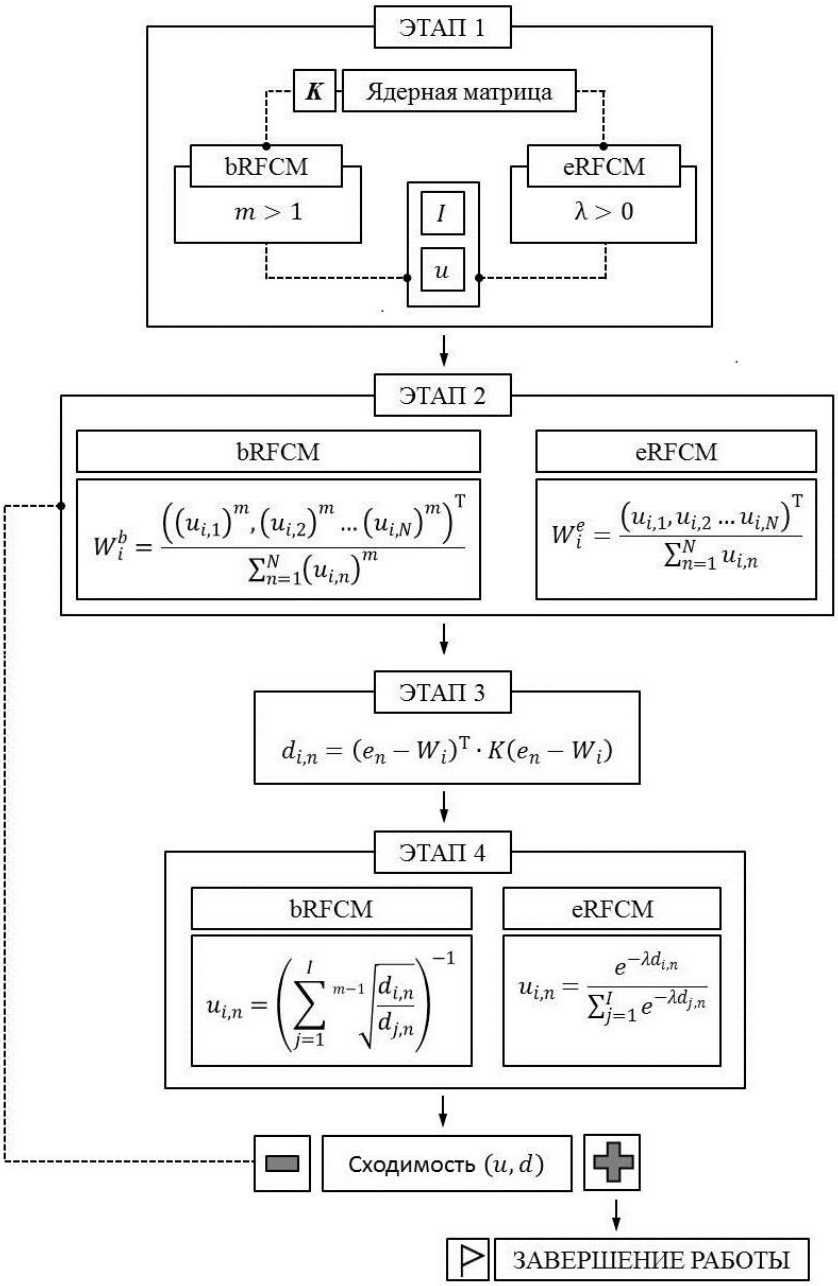

Рис. 4. Алгоритм кластеризации на основе методов K-bFCM и K-eFCM

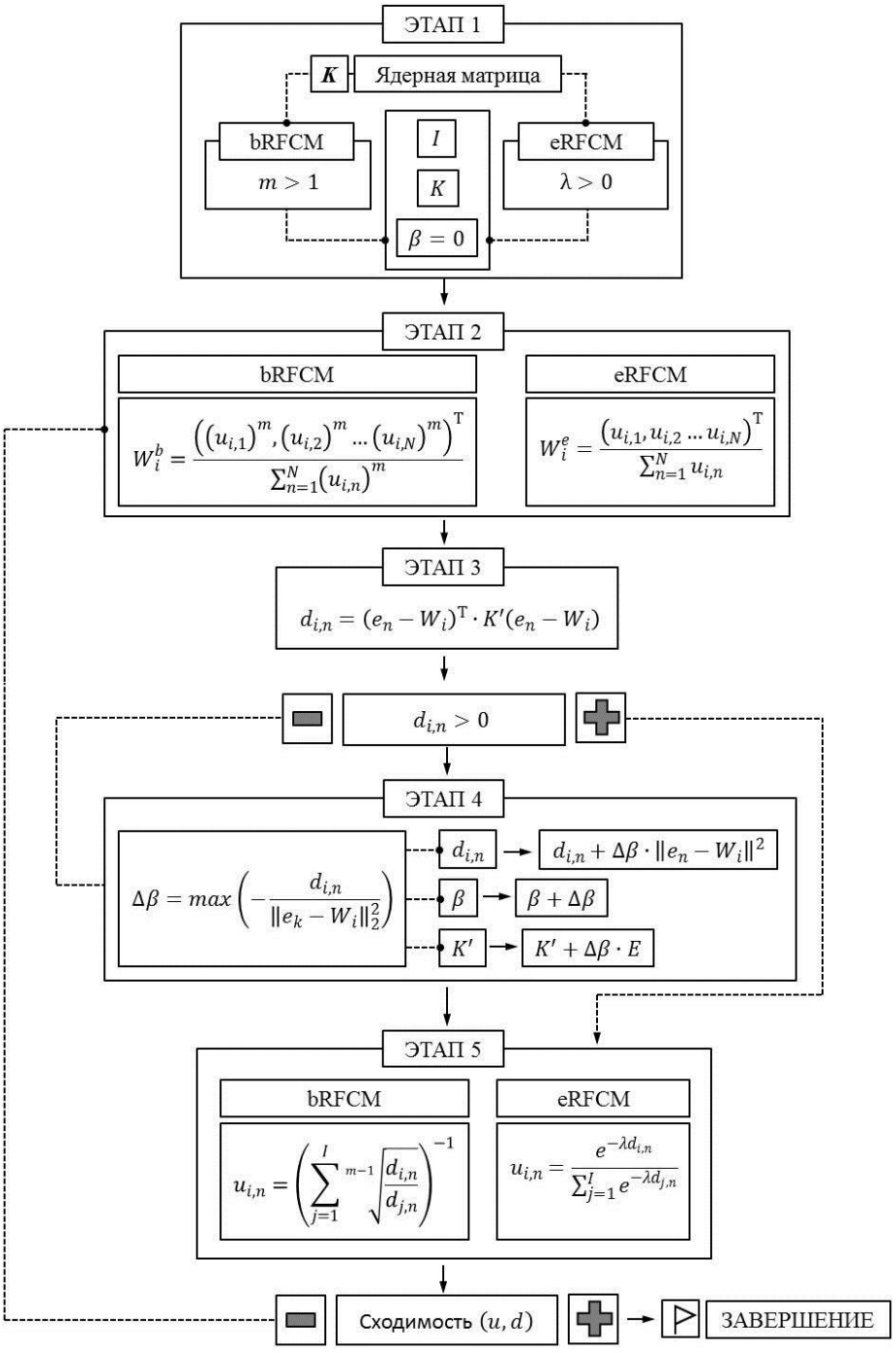

Рис. 5. Алгоритм кластеризации на основе методов K-bFCM и K-eFCM с бета-распределением

Разработанные модели позволяют решить широкий класс задач по эффективной кластеризации наборов данных методом нечетких *c*-средних для евклидовой и неевклидовой метрики.

4. Выводы

В результате проведенного анализа были предложены алгоритмы, совмещающие методы полуавтоматической кластеризации нечетких *c*-средних, в частности:

1. базовый алгоритм полуавтоматической кластеризации на основе методов нечетких c-средних Бездека;
2. алгоритм кластеризации на основе методов bRFCM и eRFCM;
3. алгоритм кластеризации на основе методов bRFCM и eRFCM для неевклидовой метрики.
4. алгоритм кластеризации на основе методов K-bFCM и K-eFCM.
5. алгоритм кластеризации на основе методов K-bFCM и K-eFCM с бета-распределением.

Список литературы / References

1. *Lee S., Kim J. & Jeong Y.*, 2017. Various Validity Indices for Fuzzy K-means Clustering. Korean Management Review. 46 (4), 1201-1226. doi:10.17287/kmr.2017.46.4.1201.
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. doi:10.1108/md-11-2015-0512.
3. *Lee J. & Lee J.*, 2014. K-means clustering based SVM ensemble methods for imbalanced data problem. 2014 Joint 7th International Conference on Soft Computing and Intelligent Systems (SCIS) and 15th International Symposium on Advanced Intelligent Systems (ISIS). doi:10.1109/scis-isis.2014.7044861.
4. A New Membership Function on Hexagonal Fuzzy Numbers. (2015). International Journal of Science and Research (IJSR), 5(5), 1129-1131. doi:10.21275/v5i5.nov163626.
5. *Miyamoto S.H., Ichihashi H. and Honda K.* Algorithms for Fuzzy Clustering, Springer, 2008.
6. *Miyamoto S. and Umayahara K.* "Fuzzy clustering by quadratic regularization," Proc. 1998 IEEE Int. Conf. Fuzzy Systems and IEEE World Congr. Computational Intelligence. Vol. 2. Pp. 1394–1399, 1998.
7. *Lewis R.H., Páláncz 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. doi:10.1145/2815111.2815138.
8. *Honda K., Oshio S. and Notsu A.* "Fuzzy co-clustering induced by multinomial mixture models," Journal of Advanced Computational Intelligence and Intelligent Informatics. Vol. 19. № 6. Pp. 717–726, 2015.
9. *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. doi:10.1002/dac.3112.
10. *Raveendran R. & Huang B.*, 2016. Mixture Probabilistic PCA for Process Monitoring - Collapsed Variational Bayesian Approach. IFAC-PapersOnLine, 49(7), 1032-1037. doi:10.1016/j.ifacol.2016.07.338.
11. *Miyamoto S. and Umayahara K.:* "Methods in Hard and Fuzzy Clustering," in: Liu, Z.-Q. and Miyamoto, S. (eds), Soft Computing and Human-centered Machines, Springer-Verlag Tokyo, 2000.
12. *Graves D. & Pedrycz W.*, 2010. Kernel-based fuzzy clustering and fuzzy clustering: A comparative experimental study. Fuzzy Sets and Systems, 161(4), 522-543. doi:10.1016/j.fss.2009.10.021.
13. *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. doi:10.1117/12.421178.
14. *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. doi:10.1109/fuzzy.2005.1452429.
15. *Kanzawa Y.:* "Entropy-Regularized Fuzzy Clustering for Non-Euclidean Relational Data and Indefinite Kernel Data," JACIII. Vol. 16, № 7. Pp. 784–792, 2012.
16. *Miyamoto S. and Suizu D.*: "Fuzzyc-Means Clustering Using Kernel Functions in Support Vector Machines," JACIII, Vol. 7, No. 1, pp. 25–30, 2003.
17. *Miyamoto S., Kawasaki Y. and Sawazaki K.:* "An Explicit Mapping for Kernel Data Analysis and Application to Text Analysis," Proc. IFSA-EUSFLAT 2009, Pp. 618–623, 2009.
18. *Kanzawa Y., Endo Y. and Miyamoto S.*: "Indefinite Kernel Fuzzyc-Means Clustering Algorithms," Lecture Notes in Computer Science, Vol. 6408, Pp. 116–128, 2010.
19. *Bouchachia A. and Pedrycz W.:* "Data Clustering with Partial Supervision," Data Mining and Knowledge Discovery. Vol. 12. Pp. 47–78, 2006.
20. *Yamazaki M., Miyamoto S. and Lee I.-J.:* "Semi-supervised Clustering with Two Types of Additional Functions," Proc. 24th Fuzzy System Symposium. 2E2-01, 2009.
21. *Macario V. & Francisco De A.T. De Carvalho*, 2010. A new approach for semi-supervised clustering based on Fuzzy C-Means. International Conference on Fuzzy Systems. doi:10.1109/fuzzy.2010.5584306.
22. *Yamashiro M., Endo Y., Hamasuna Y. and Miyamoto S.:* "A Study on Semi-supervised Fuzzy c-Means," Proc. 24th Fuzzy System Symposium, 2E3-04, 2009.
23. *Kanzawa Y., Endo Y. and Miyamoto S.:* "A Semi-Supervised Entropy Regularized Fuzzy c-Means," Proc. 2009 International Symposium on Nonlinear Theory and Its Applications, Pp. 564–567, 2009.
24. *Liu L. & Wu X.*, 2013. Semi-Supervised Possibilistic Fuzzy c-Means Clustering Algorithm on Maximized Central Distance. Proceedings of the 2nd International Conference on Computer Science and Electronics Engineering (ICCSEE 2013). doi:10.2991/iccsee.2013.342.
25. *Kanzawa Y., Endo Y. and Miyamoto S:* "Some Pairwise Constrained Semi-Supervised Fuzzy c-Means Clustering," LNAI, Vol. 5681, Pp. 268–281, 2009.
26. *Thong P.H. & Son L.H.*, 2016. An Overview of Semi-Supervised Fuzzy Clustering Algorithms. International Journal of Engineering and Technology. 8 (4), 301-306. doi:10.7763/ijet.2016.v6.902.
27. *Kanzawa Y., Endo Y. and Miyamoto S.:* "Semi-Supervised Fuzzy c-Means Algorithm by Revising Dissimilarity Between Data," JACIII. Vol. 15, № 1. Pp. 95–101, 2011.