ПРИМЕНЕНИЕ ПАРАДИГМЫ НЕЧЕТКОЙ КЛАСТЕРИЗАЦИИ И БИКЛАСТЕРИЗАЦИИ ПРИ МОНИТОРИНГЕ ИНФРАСТРУКТУРЫ ЦЕНТРОВ ОБРАБОТКИ ДАННЫХ

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 нечеткой кластеризации, в частности применение метода нечетких с-средних. Показана необходимость построения теоретической методологии использования метода нечетких с-средних. Рассмотрены модели нечеткой кластеризации, которые базируются на концепции смесей вероятностных распределений, а также введении в статистическую модель алгоритмов нечеткой регулируемой коррекции. При этом метод нечетких с-средних, основанный на энтропийной регуляризации, рассматривается в рамках модели смеси гауссовых распределений и фаззификации, сравнивается по эффективности с классическим методом нечетких с-средних. Помимо этого, концепция регуляризации обсуждается в контексте нечеткой бикластеризации, а также рассматривается полиноминальная модель кластеризации. На основе результатов экспериментальной верификации данных моделей показано, что модель нечеткой кластеризации, которая базируется на концепции смесей вероятностных распределений и введении в статистическую модель алгоритмов нечеткой регулируемой коррекции демонстрирует улучшение интерпретируемости разбиения объекта на кластеры.*

Ключевые слова: *информационные системы, метод нечетких с-средних, метод энтропийной регуляризации, смеси гауссовых распределений, фаззификация, нечеткая бикластеризация, полиноминальная модель кластеризации.*

УДК 331.225.3

Введение: Внедрение метода кластерного анализа данных путем построения групп объектов информационных систем на основании ключевых параметров, определяющих их сходство, широко используется в современных информационных технологиях [1-5], в частности при мониторинге и классификации объектов центров обработки данных, что указывает на ***актуальность разработки*** теоретической методологии использования данного подхода.

Анализ последних исследований и публикаций в данной области показал перспективность применения иерархических алгоритмов, в то время как неиерархические алгоритмы продолжают использоваться в мобильных приложениях, поскольку требуют меньшие вычислительные мощности. Так метод на основе нечетких *k*-средних [2, 6] является самым известным неиерархическим алгоритмом кластеризации, хотя на сегодняшний день в большей степени используются модернизированные алгоритмы на его основе. В то же время алгоритмы на основе метода нечетких *c*-средних применяются для проведения нечеткого разбиения [1, 3, 7], через внедрение парадигмы нечеткого набора, где нечеткое разбиение реализуется путем введения целевой функции нелинейного характера с весовым показателем.

Другая модель фаззификации базируется на парадигме регуляризации энтропии, в которой дополнительный нелинейный член (который обычно является квадратичным) комбинируется с целевой функцией *k*-средних [8, 9]. Было показано, что алгоритм на основе кластеризации типа *k*-средних также может иметь другую интерпретацию с точки зрения парадигмы смеси вероятностных распределений. Таким образом, функция правдоподобия (likelihood function) смеси гауссовых распределений (GMM: Gaussian Mixture Models) может быть разложена на целевую функцию с жестким *k*-средним и дополнительный член на основе мягкого разбиения [10-12]. Данная парадигма поддерживает достоверность энтропии регуляризованной целевой функции метода нечетких *c*-средних и подразумевает связь кластеризацией метода нечетких *c*-средних и моделей смесей вероятностных распределений. При этом были предложены нечеткие аналоги нескольких вероятностных моделей смесей, где степень нечеткости вероятностных разбиений настраивается с весовых коэффициентов [13, 14].

Кластеризация на основе метода нечетких *c*-средних также может быть расширена до нечеткой бикластеризации (fuzzy co-clustering), где цель состоит в том, чтобы извлечь парные кластеры объектов набора на основе информации о совпадениях. Помимо моделей регуляризации на основе энтропии [15-17], был предложен нечеткий аналог моделей полиномиальной смеси [18, 19], который также реализуется с регулируемым весовым коэффициентом.

Целью работы, таким образом, стала разработка комплексной методологии использования алгоритмов на основе метода нечетких *c*-средних и *k*-средних в центрах обработки данных путем обобщения приведенных выше моделей и проведения численных экспериментов

1. Внедрение фаззификации в методах нечетких c-средних и k-средних

Фаззификация или введение нечеткости является процессом установления соответствия между численным значением входных данных нечеткого вывода и значением функции принадлежности соответствующего ей терма. Таким образом, при фаззификации в соответствие значениям всех входных данных системы ставятся конкретные значения функций принадлежности соответствующих термов. Цель кластеризации типа *k*-средних состоит в том, чтобы разделить объекты x_i, где $i \in [1;I]$, на кластеры $c \in [1;C]$ множеств с репрезентативными центроидами b_c, где внутрикластерные объекты максимально похожи друг на друга. Соответственно, алгоритм кластеризации *k*-средних включает случайное назначение центроидов и оптимизацию центроидов через разбиение элементов системы до уровня сходимости. Алгоритмы разделения могут быть представлены несколькими моделями функции принадлежности (membership function) в соответствии с различными ограничениями.

Пусть есть группа объектов инфраструктуры центра обработки данных общим числом $i \in [1;I]$, которая может быть представлена кластерами $c \in [1;C]$. В таком случае ограничения для четкого с-разделения u_{ci}^{H} (hard *c*-partition), нечеткого с-разделения u_{ci}^{F} (fuzzy *c*-partition) и вероятностного с-разделения u_{ci}^{P} (possibilistic *c*-partition), где u_{ci} — функция принадлежности, могут быть определены как:

$$\left[\begin{cases} u_{ci}^{H} \in \{0;1\} \\ \forall\, i \in [1;I] \\ \forall\, c \in [1;C] \end{cases} \begin{cases} \sum_{c=1}^{C} u_{ci}^{H} = 1 ; \\ \forall\, i \in [1;I] \end{cases} \begin{cases} \sum_{i=1}^{I} u_{ci}^{H} > 0 \\ \forall\, c \in [1;C] \end{cases}\right. \quad \left[\begin{cases} u_{ci}^{F} \in [0;1] \\ \forall\, i \in [1;I] \\ \forall\, c \in [1;C] \end{cases} \begin{cases} \sum_{c=1}^{C} u_{ci}^{F} = 1 ; \\ \forall\, i \in [1;I] \end{cases} \begin{cases} \sum_{i=1}^{I} u_{ci}^{F} > 0 \\ \forall\, c \in [1;C] \end{cases}\right. \quad \left[\begin{cases} u_{ci}^{P} \in [0;1] \\ \forall\, i \in [1;I] \\ \forall\, c \in [1;C] \end{cases} \begin{cases} \sum_{i=1}^{I} u_{ci}^{P} > 0 \\ \forall\, c \in [1;C] \end{cases}\right. . \qquad (1)$$

Как можно видеть, при переходе $u_{ci}^{H} \to u_{ci}^{F} \to u_{ci}^{P}$ жесткость ограничений уменьшается, и критерии назначения объектов набора становится гораздо более гибким.

Алгоритмы на основе метода нечетких *c*-средних строятся на основе нечеткого *c*-разделения и модифицированной целевой функции *k*-средних. Стандартная модель гибридной нечеткой кластеризации также включает дополнительный весовой показатель *m*, где $m > 1$ (рис. 1).

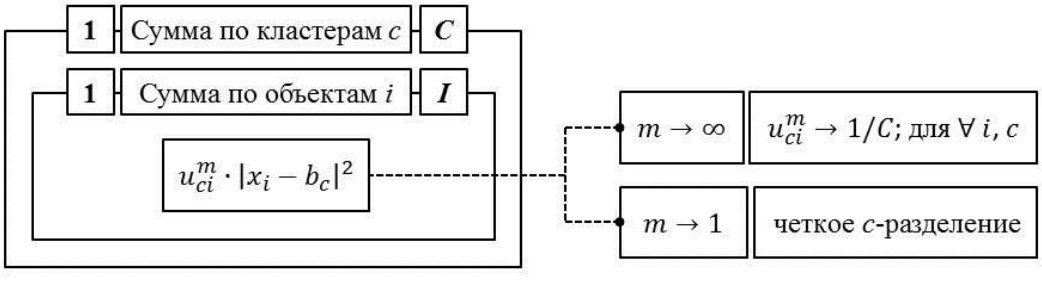

Рис. 1. Схема определения значения функции правдоподобия по методу гибридной нечеткой кластеризации

Как показано на рис. 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 $u_{ci}^{m} \to 1/C$ для любых объектов и кластеров объектов.

В свою очередь метод энтропийной нечеткой кластеризации (рис. 2) объединяет определение целевой функции по методу k-средних с энтропийной штрафной функцией (entropy-like penalty function).

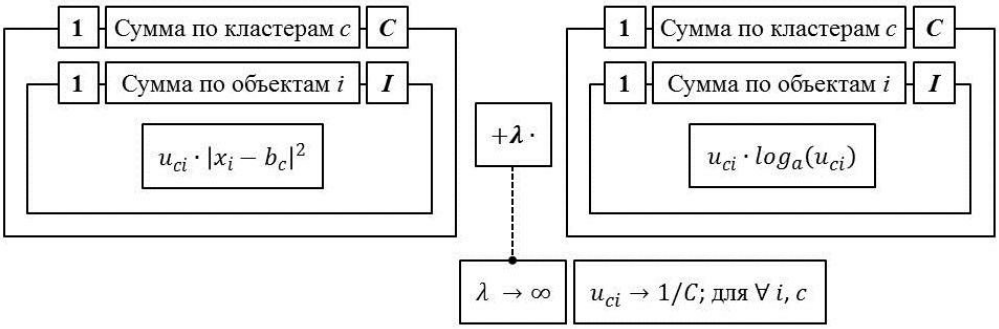

Рис. 2. Схема определения значения функции правдоподобия по методу энтропийной нечеткой кластеризации

Соответственно для метода энтропийной нечеткой кластеризации при росте λ значение $u_{ci}^{m} \to 1/C$ для любых объектов и кластеров объектов.

2. Нечеткая кластеризация для смесей вероятностных распределений

Ограничения нечеткой кластеризации представленные в системах уравнений (1) могут быть соотнесены с парадигмой смесей вероятностных распределений [20-22] через соотнесение функция принадлежности u_{ci} и порождающей вероятности (generative probability) объекта x_i, где $i \in [1; I]$, который относится к -распределению. Пусть объекты взяты из одного из независимых гауссовых распределений, тогда каждый из них представляет собой гауссов компонент g_c со средним b_c и корреляционным моментом cov_c. Таким образом, вероятность может быть рассчитана через g_c и весовой коэффициент α_c:

$$P_i = \sum_{c=1}^{C} \alpha_c \cdot g_c(x_i | b_c, cov_c). \qquad (2)$$

В свою очередь на основе расчета вероятности для смеси гауссовых распределений может быть определена функция правдоподобия:

$$L_G = \sum_{i=1}^{I} log(P_i). \qquad (3)$$

При помощи неравенства Йенсена может быть найдено решение для максимизации функции правдоподобия:

$$L_{GJ=} \sum_{c=1}^{C} \sum_{i=1}^{I} u_{ci} \cdot log\left(\frac{\alpha_i \cdot g_c(x_i | b_c, cov_c)}{u_{ci}}\right). \qquad (4)$$

На основе данного выражения может быть построен алгоритм, представленный на рис. 3.

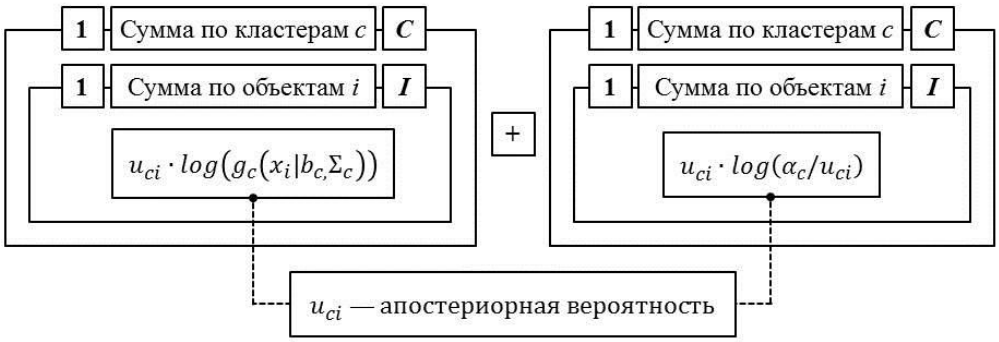

Рис. 3. Схема определения значения функции правдоподобия для смеси гауссовых распределений

При этом u_{ci} является апостериорной вероятностью для гауссова компонента c с параметрами $\{x_i, b_c, cov_c\}$, объекта x_i и функции принадлежности u_{ci}.

В рамках данной работы предлагается использовать нечеткий аналог полнопараметрических смеси гауссовых распределений путем применения расхождения Кульбака-Лейблера (РКЛ). Максимум целевой функции при этом может быть получен через расширения метода нечетких *c*-средних:

$$\begin{cases} L_{\text{РКЛ}} = \sum_{c=1}^{C}\sum_{i=1}^{I} u_{ci}\,(\Delta_i)^{\mathrm{T}} cov_c^{-1}(\Delta_i) - \\ -\lambda\sum_{c=1}^{C}\sum_{i=1}^{I} u_{ci}\log\frac{\alpha_c}{u_{ci}} + \sum_{c=1}^{C} u_{ci}\log|cov_c| \\ \alpha_c = \frac{1}{I}\sum_{i=1}^{I} u_{ci} \\ b_c = \frac{\sum_{i=1}^{I} u_{ci}x_i}{\sum_{i=1}^{I} u_{ci}} \\ cov_c = \frac{1}{\sum_{i=1}^{I} u_{ci}}\sum_{i=1}^{I}(\Delta_i \cdot (\Delta_i)^{\mathrm{T}}) \\ u_{ci} = \frac{\alpha_c exp(-d_{ci})}{\sum_{l=1}^{C}\alpha_l exp(-d_{ci})} \\ \begin{bmatrix} d_{ci} = (\Delta_i)^{\mathrm{T}}\Sigma_c(\Delta_i) \\ \Delta_i = x_i - b_c \end{bmatrix} \end{cases}, \qquad (5)$$

где: λ — регулируемый вес для настройки степени нечеткости разделения, чем больше значение λ, тем более нечетким является разделение.

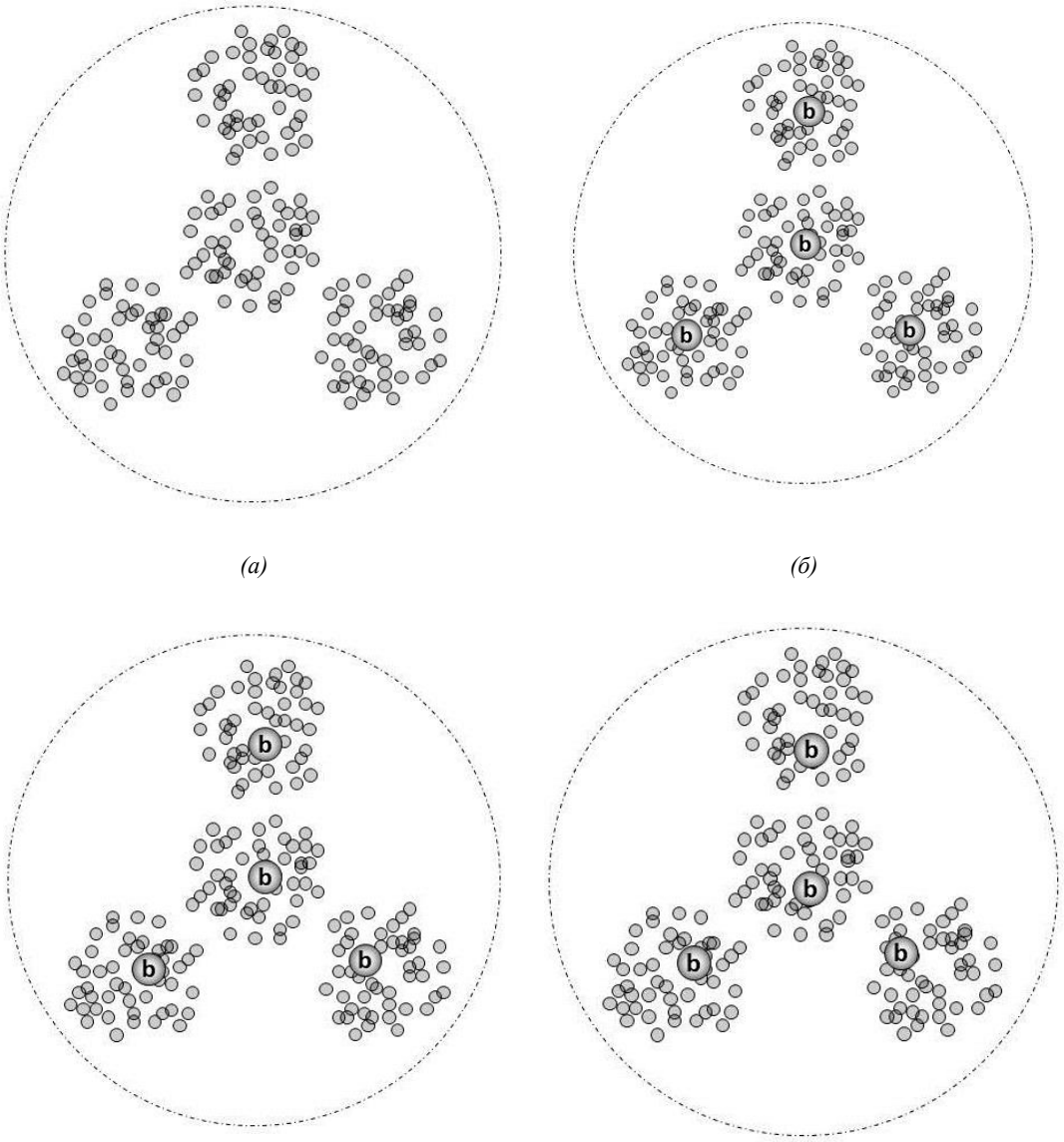

Рис. 4. Численное моделирование кластерного РКЛ-разделения: (а) набор объектов в двумерном пространстве; (б) разделение для $\lambda = 0{,}5$*; (в) разделение для* $\lambda = 1$*; (г) разделение для* $\lambda = 2$

Для определения эффективности алгоритма определения максимум целевой функции через расширения метода нечетких *c*-средних на основе РКЛ анализа было проведено численное моделирование. Предложенная модель включает в себя набор объектов представленных в двумерном пространстве, на основе которого можно образовать четыре кластера с равным количеством объектов в каждом.

На рис. 4 представлены результаты численного моделирования кластерного РКЛ-разделения для $C = 4$ и $\lambda = 0{,}5$ (рис. 4-б), $\lambda = 1$ (рис. 4-в), $\lambda = 2$ (рис. 4-г). Результат, полученный для $\lambda = 1$, демонстрирует влияние на расположение центроида центрального кластера элементов внешних кластеров, таким образом, можно видеть, что при нечетком разделении могут искажаться границы кластеров, неоднозначность которых связана с определением негауссовых плотностей компонентов. Соответственно результат полученный для $\lambda = 0{,}5$ указывает на более четкое разделение с $\lambda = 0{,}5$, что является предпочтительным для уточнения границ кластеров. Следует отметить, что применение более четкой модели, дает возможность воспользоваться преимуществами как четкого разбиения *k*-средних,

так и нечеткого определения принадлежности элементов набора данных. С другой стороны, при том, что для $\lambda = 2$ на центральный кластер оказывается еще больше влияние внешних кластеров, центр кластера корректно отображается в его центроиде.

Проведенное моделирование показывает, что регулирование параметров кластерного анализа может способствовать улучшению интерпретируемости результата разделение набора на кластеры. Более четкая модель подходит для линейно разделяемых наборов данных, в то время как нечеткая модель эффективно работает при анализе перекрывающихся кластеров.

Выводы

В результате проведенного анализа были изучены современные подходы нечеткой кластеризации, в частности применение метода нечетких *c*-средних для смесей вероятностных распределений, и сделаны выводы по их применения в информационных системах, в частности:

1. Проведен анализ ограничений для четкого с-разделения, нечеткого с-разделения и вероятностного с-разделения. Математический аппарат, который применяется в данных моделях, был соотнесен с уравнениями вероятностного анализа.
2. Был рассмотрен подход на основе применения парадигмы смесей вероятностных распределений, в рамках которого объекты, каждый из них представляет собой гауссов компонент, могут быть взяты из одного из независимых гауссовых распределений, а вероятность рассчитывается через весовой коэффициент.
3. Был предложен алгоритм по использованию полнопараметрической смеси гауссовых распределений путем применения расхождения Кульбака-Лейблера, где максимум целевой функции при этом может быть получен через расширения метода нечетких *c*-средних.
4. Было проведено численное моделирование и показано, что регулирование параметров кластерного анализа способствует улучшению интерпретируемости результата разделение набора на кластеры.

Список литературы

1. *Haqiqi B.N. & Kurniawan R.*, 2015. Analisis Perbandingan Metode Fuzzy C-Means Dan Subtractive Fuzzy C-Means. Media Statistika, 8 (2). doi:10.14710/medstat.8.2.59-67.
2. *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.
3. *Yasuda M.*, 2014. Q-increment deterministic annealing fuzzy c-means clustering using Tsallis entropy. 2014 11th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD). doi:10.1109/fskd.2014.6980802.
4. *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.
5. *Baili N.*, 2013. Unsupervised and semi-supervised fuzzy clustering with multiple kernels. Louisville, KY: University of Louisville.
6. *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.
7. 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.
8. *Miyamoto S., Ichihashi H. and Honda K.* Algorithms for Fuzzy Clustering. Springer, 2008.
9. *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.
10. *Bishop C.M.* Neural Networks for Pattern Recognition, Clarendon Press, 1995.
11. *Hualde J. & Robinson P.M.*, 2011. Gaussian pseudo-maximum likelihood estimation of fractional time series models. The Annals of Statistics, 39(6), 3152-3181. doi:10.1214/11-aos931.
12. *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. doi:10.1145/2815111.2815138.
13. *Ichihashi H., Miyagishi K. and Honda K.* "Fuzzyc-means clustering with regularization by K-L information", Proc. of 10th IEEE International Conference on Fuzzy Systems, Vol.2, Pp. 924–927, 2001.
14. *Honda K. and Ichihashi H.* "Regularized linear fuzzy clustering and probabilistic PCA mixture models", IEEE Trans. Fuzzy Systems. Vol. 13. № 4. Pp. 508–516, 2005.
15. *Ichihashi H., Notsu A. & Honda K.*, 2010. Semi-hard c-means clustering with application to classifier design. International Conference on Fuzzy Systems. doi:10.1109/fuzzy.2010.5584553

16. *Oh C.-H., Honda K. and Ichihashi H.* "Fuzzy clustering for categorical multivariate data," Proc. of Joint 9th IFSA World Congress and 20th NAFIPS International Conference. Pp. 2154–2159, 2001.
17. *Kummamuru K., Dhawale A. and Krishnapuram R.* "Fuzzy co-clustering of documents and keywords," Proc. 2003 IEEE Int'l Conf. Fuzzy Systems. Vol. 2. Pp. 772–777, 2003.
18. *Rigouste L., Cappé O.and Yvon F.* "Inference and evaluation of the multinomial mixture model for text clustering," Information Processing and Management, Vol. 43, no. 5, Pp. 1260–1280, 2007.
19. *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, no. 6, pp. 717–726, 2015.
20. *Kumar P. & Chaturvedi A.,* 2016. Probabilistic query generation and fuzzyc-means clustering for energy-efficient operation in wireless sensor networks. International Journal of Communication Systems, 29(8), 1439-1450. doi:10.1002/dac.3112.
21. *Wang Z., Wang L., Dang H. & Pan L.,* 2013. Web clustering based on hybrid probabilistic latent semantic analysis model. Journal of Computer Applications, 32 (11), 3018-3022. doi:10.3724/sp.j.1087.2012.03018.
22. *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.