УДК 519.6:004.93

Г.Ю. Щербакова, О.В. Логвинов, кандидаты техн. наук**, В.Н. Крылов,** д-р техн. наук

НЕЧЕТКАЯ КЛАСТЕРИЗАЦИЯ В ПРОСТРАНСТВЕ ВЕЙВЛЕТ - ПРЕОБРАЗОВАНИЯ

Разработан субградиентный метод нечеткой кластеризации. Применение метода позволит повысить помехоустойчивость и вероятность выхода в область глобального оптимума при кластеризации при современном автоматизированном производственном контроле.

***Ключевые слова**: нечеткая кластеризация, вейвлет - преобразование, субградиент.*

G.Y.Shcherbakova, Ph.D., **V. N.Krylov,** Ph.D., **O. V Logvinov,** ScD.

FUZZY CLUSTERING IN THE WAVELET TRANSFORMING DOMAIN

The sub gradient fuzzy clustering method is designed. Such method allows noise stability and possibility of the global optimum achievement rising for the clustering procedure in time of modern plant automated control.

***Keywords:** Fuzzy clustering, wavelet transforming, sub gradient.*

Г.Ю. Щербакова, О.В. Логвінов, кандидати техн.наук, **В.М .Крилов,** д-р техн. наук

НЕЧІТКА КЛАСТЕРИЗАЦІЯ В ПРОСТОРІ ВЕЙВЛЕТ – ПЕРЕТВОРЕННЯ

Розроблено субградієнтний метод нечіткої кластеризації. Його застосування дає змогу підвищити завадостійкість і ймовірність виходу в область глобального оптимуму при кластеризації у разі автоматизованого контролю на сучасному виробництві.

***Ключові слова**: нечітка кластеризація, вейвлет - перетворення, субградієнт.*

Отладка технологических процессов (ТП) современного производства проводится на основе автоматизированного контроля параметров объектов производства.

Состояние объектов производства при этом описывается значением их параметров $X^n \in S_p^{(n)}$. Здесь $X^n = \{X_1, X_2, \ldots, X_n\}$ - модель объекта производства, представленная n - мерным случайным вектором; $S_p^{(n)}$ - n -мерная допусковая область. В случае, если среди n параметров не найден определяющий, оперативный контроль реализуется автоматизированными системами технического диагностирования (АСТД) путем анализа многомерных массивов параметров посредством классификации при распознавании образов [5]. При таком контроле в группе объектов производства выделяются компактные подгруппы (кластеры) и оценивается их близость к границе поля допуска.

Поскольку на стадии отладки ТП еще не известно, к какому кластеру относится та или иная точка в признаковом пространстве (образ), но необходимо определить границы между кластерами, АСТД должны реализовывать классификацию с самообучением, которая включает две процедуры: кластеризацию и классификацию.

С позиций теории управления большинство ТП представляют собой нелинейный динамический нестационарный стохастический объект, характеризующийся высоким уровнем априорной неопределенности. В условиях современного мелкосерийного производства уровень этой неопределенности повышается, так как АСТД должны обеспечивать заданную достоверность при контроле по параметрам малых выборок изделий и оперативность перестройки и дополнительного обучения из-за частой смены их номенклатуры.

При кластеризации эта неопределенность проявляется в невозможности адекватно оценить по параметрам малой выборки плотность вероятности, характеризующую принадлежность объекта кластеру. Кроме того, кластеры могут быть линейно не сепарабельными, иметь сложную форму, значительно различаться по количеству значений и локальной плотности расположения точек в признаковом пространстве.

Это обусловливает необходимость проведения нечеткой кластеризации, формирующей приближение к такой оценке путем

оценки взвешенной принадлежности объекта кластеру и позволяющей более адекватно представлять объекты, находящиеся на границах кластеров [1].

Основные методы нечеткой кластеризации – иерархические и итеративные – отличаются рядом недостатков. Основной недостаток иерархических методов в зависимости от принятой меры расстояния – низкая помехоустойчивость.

Итеративные методы, оптимизируя некоторый функционал качества, разделяют объекты на группы (кластеры) с учетом сходства, определяемого через расстояние либо между объектами, либо от объектов к центру кластера. По оценке расстояния функционалы разделяют на две группы.

Параметры функционалов первой группы, основанных на определении расстояния между объектами, часто определить сложно, а результаты кластеризации зависят от локальной плотности распределения данных, поэтому чаще применяют функционалы второй группы – сформированные путем обобщения функционала *c*- средних [1, 2, 6, 9-11]. Эти функционалы сформированы для снижения чувствительности результатов кластеризации к помехам в данных либо к начальным значениям центров кластеров [9, 10].

Добиваются этих целей либо меняя метрику пространства признаков, в котором проводится кластеризация, (вместо евклидова расстояния, например, применяя расстояние Махалонобиса), либо добавляя к базовому функционалу c- средних слагаемые для учета локальной плотности распределения данных [9, 10], либо вводя промежуточные этапы обработки данных. Кластеризация с помощью таких проблемно-ориентированных функционалов позволяет снизить влияние только одного из указанных факторов с увеличением времени обработки данных почти на порядок [9, 10]. Поиск оптимума этих функционалов качества, которые при нечеткой кластеризации могут обладать многоэкстремальной поверхностью [6], как правило, реализуют, применяя аппарат нелинейного программирования, основанный на множителях Лагранжа [2, 9].

Процедура поиска оптимума, помимо указанного выше, усложняется тем, что запись правила этих множителей имеет не единственное решение и сходимость метода требует хорошего начального приближения [4]. В таких условиях основные недостатки итеративных методов оптимизации, на которых основана нечеткая кластеризация, – чувствительность к начальной точке поиска и шуму в данных, отыскание не глобального, а локального минимума – усугубляются особенностями оценки градиента или субградиента, используемых при их реализации. Поэтому методы нечеткой кластеризации на основе градиентного поиска отличаются низкой помехоустойчивостью, а субградиентные методы - низкой точностью. В связи с этим в ряде приложений оптимизацию при нечеткой кластеризации проводят на основе генетических алгоритмов и реализуют, распараллеливая вычисления и увеличивая количество процессоров [6]. Такой подход значительно снижает оперативность при техническом диагностировании в процессе отладки технологических процессов производства.

Для оптимизации в указанных условиях авторами разработан субградиентный метод оптимизации с повышенной помехоустойчивостью, пониженными погрешностью, чувствительностью к локальным экстремумам и начальной точке поиска [3].

Цель работы – разработка на основе этого метода субградиентного итеративного метода нечеткой кластеризации в пространстве вейвлет - преобразования (ВП) для повышения помехоустойчивости и повышения вероятности выхода в область глобального оптимума.

Указанный метод разработан на базе одного из наиболее общих [2] итеративных методов кластеризации, связанного с минимизацией среднего риска, и основан на требовании, чтобы элемент кластера был удален от его центра на расстояние меньшее, чем от центров $\mathbf{c}_{k-1}$ остальных $k-1$ кластеров. Для евклидовых расстояний функция потерь k-го кластера будет $F_k(\mathbf{x}, \mathbf{c}_k) = (\mathbf{x} - \mathbf{c}_k)^2$, а функционал – эквивалентен реализации функционала кластеризации k-средних [2]. Взвешенный вариант этого функционала – функционал c– средних оценивает нечеткую близость объектов к центрам кластеров [2, 7]

$$Q(\mathbf{x}, \mathbf{c}) = \sum_{k=1}^{M}\sum_{i=1}^{N} \mu_{k,i}^{\beta} D_{k,i}$$

при ограничениях $\sum_{k=1}^{M} \mu_{k,i} = 1, \quad \mu_{k,i} \in [0,1]$, $1 \le i \le N$. Здесь $\mu_{k,i}$ – функции принадлежности i-го объекта кластеру k; β – параметр нечеткости в разбиении объектов на кластеры (в работе $\beta = 2$); $D_{k,i} = \sqrt{\sum_{l=1}^{L} (x_{i,l} - c_{k,l})^2}$ –расстояния между объектами и центрами кластеров L – размерность пространства признаков.

Минимизация функционала реализуется на основе итераций Пикара [7], когда после инициализации параметров алгоритма количества кластеров M, параметров нечеткости β и останова ε, генерируется начальная матрица нечеткого разбиения $\mathrm{M} = [\mu_{k,i}]$ и последовательно рассчитываются c_k и $\mu_{k,i}$ пока не выполнится условие останова

$$c_k = \frac{\sum_{i=1}^{N} \mu_{k,i}^{\beta} \cdot x_{i,l}}{\sum_{i=1}^{N} \mu_{k,i}^{\beta}}$$

$$\mu_{k,i} = \frac{1}{\left(D_{k,l}^2 \sum_{p=1}^{M} \frac{1}{D_{p,i}^2} \right)^{\frac{1}{\beta-1}}}, \qquad (1)$$

Подобный подход положен в основу разработанного метода кластеризации. Для минимизации среднего риска при кластеризации определяют оптимальный вектор $\mathbf{c} = \mathbf{c}_{opt}$, который, удовлетворяя ограничениям, доставлял бы экстремальное значение $Q(\mathbf{x}, \mathbf{c})$ – функционалу вектора переменных $\mathbf{c} = (c_1, \ldots, c_N)$, зависящему от вектора случайных последовательностей $\mathbf{x} = (x_1, \ldots, x_M)$. По показам образов $x \in X$ определяют центры X_k и границы множеств. При этом $\mathrm{F_k}(\mathbf{x}, \mathbf{c}_1, \ldots, \mathbf{c}_\mathrm{M})$ - функция расстояния элементов $\mathbf{x}$ множества X от центров кластеров $\mathbf{c}_\mathrm{k}$; реализация функционала качества

$$Q(\mathbf{x}, \mathbf{c}_1, \ldots, \mathbf{c}_\mathrm{M}) = \sum_{k=1}^{M} \mu_k(\mathbf{x}, \mathbf{c}_1, \ldots, \mathbf{c}_\mathrm{M}) \mathrm{F_k}(\mathbf{x}, \mathbf{c}_1, \ldots, \mathbf{c}_\mathrm{M}).$$

Для двух классов функция потерь

$$S(\mathbf{x}, \mathbf{c}_1, \mathbf{c}_2) = \mu_1(\mathbf{x}, \mathbf{c}_1, \mathbf{c}_2) F_1(\mathbf{x}, \mathbf{c}_1, \mathbf{c}_2) + \mu_2(\mathbf{x}, \mathbf{c}_1, \mathbf{c}_2) F_1(\mathbf{x}, \mathbf{c}_1, \mathbf{c}_2),$$

а условия минимума среднего риска

$$\left.\begin{aligned} M_x\{\mu_1(\mathbf{x}, \mathbf{c}_1, \mathbf{c}_2) \widetilde{\nabla}_{c1\pm} F_1(\mathbf{x}, \mathbf{c}_1, \mathbf{c}_2) + \\ + \mu_2(\mathbf{x}, \mathbf{c}_1, \mathbf{c}_2) \widetilde{\nabla}_{c1\pm} F_2(\mathbf{x}, \mathbf{c}_1, \mathbf{c}_2)\} = 0 \\ M_x\{\mu_1(\mathbf{x}, \mathbf{c}_1, \mathbf{c}_2) \widetilde{\nabla}_{c2\pm} F_1(\mathbf{x}, \mathbf{c}_1, \mathbf{c}_2) + \\ + \mu_2(\mathbf{x}, \mathbf{c}_1, \mathbf{c}_2) \widetilde{\nabla}_{c2\pm} F_2(\mathbf{x}, \mathbf{c}_1, \mathbf{c}_2)\} = 0 \end{aligned}\right\}.$$

Тогда алгоритм нечеткой кластеризации

$$\left.\begin{aligned} \mathrm{c}_1[n] = \\ = c_1[n-1] - \gamma_1[n] \mu_1(\mathbf{x}[\mathrm{n}], \mathbf{c}[n-1]) \times \\ \times \sum_{\mathrm{m}=1}^{s_\alpha} \alpha_\mathrm{m}[\mathrm{n}] \widetilde{\nabla}_{\mathrm{c1}_-^+} \mathrm{F_1}(\mathbf{x}[\mathrm{n}], \mathbf{c}[\mathrm{n-1}], a[\mathrm{n-m}]), \\ \mathrm{c}_2[n] = \\ = c_2[n-1] - \gamma_2[n] \mu_2(\mathbf{x}[\mathrm{n}], \mathbf{c}[n-1]) \times \\ \times \sum_{\mathrm{m}=1}^{s_\alpha} \alpha_\mathrm{m}[\mathrm{n}] \widetilde{\nabla}_{\mathrm{c2}_-^+} \mathrm{F_1}(\mathbf{x}[\mathrm{n}], \mathbf{c}[\mathrm{n-1}], a[\mathrm{n-m}]) \end{aligned}\right\} \qquad (2)$$

при

$$f(\mathbf{x}[\mathrm{n}], \mathbf{c}[n-1]) = \mathrm{F_1}(\mathbf{x}[\mathrm{n}], \mathbf{c}[n-1]) - \mathrm{F_2}(\mathbf{x}[\mathrm{n}], \mathbf{c}[n-1]) \le 0$$

и

$$\left.\begin{aligned} \mathrm{c}_1[n] = \\ = c_1[n-1] - \gamma_1[n] \mu_1(\mathbf{x}[\mathrm{n}], \mathbf{c}[n-1]) \times \\ \times \sum_{\mathrm{m}=1}^{s_\alpha} \alpha_\mathrm{m}[\mathrm{n}] \widetilde{\nabla}_{\mathrm{c1}_-^+} \mathrm{F_2}(\mathbf{x}[\mathrm{n}], \mathbf{c}[\mathrm{n-1}], a[\mathrm{n-m}]), \\ \mathrm{c}_2[n] = \\ = c_2[n-1] - \gamma_2[n] \mu_2(\mathbf{x}[\mathrm{n}], \mathbf{c}[n-1]) \times \\ \times \sum_{\mathrm{m}=1}^{s_\alpha} \alpha_\mathrm{m}[\mathrm{n}] \widetilde{\nabla}_{\mathrm{c2}_-^+} \mathrm{F_2}(\mathbf{x}[\mathrm{n}], \mathbf{c}[\mathrm{n-1}], a[\mathrm{n-m}]) \end{aligned}\right\} \qquad (3)$$

при

$$f(\mathbf{x}[\mathrm{n}], \mathbf{c}[n-1]) = \mathrm{F_1}(\mathbf{x}[\mathrm{n}], \mathbf{c}[n-1]) - \mathrm{F_2}(\mathbf{x}[\mathrm{n}], \mathbf{c}[n-1]) > 0.$$

Здесь $\gamma_k[n]$ – шаг; n – номер итерации; – оценка субградиента функции потерь для i-го кластера путем ВП реализации $\mathrm{F_i}(\mathbf{x}, \mathbf{c})$ по $c_i, i = 1, \ldots, N$; a – скаляр; $\alpha_m[n]$, $m = 1, \ldots, s_\alpha$ - компоненты вектора $\alpha[n]$ после дискретизации вейвлет – функции (ВФ).

Сначала для оценки субградиента используют свертку функционала с ВФ Хаара в окрестности, определяемой длиной ее носителя L [3]. При этом обеспечивается достаточная помехоустойчивость при небольших вычислительных затратах, но (из-за асимметрии функционала) погрешность поиска экстремума на этом этапе высока. Для снижения этой погрешности на следующем этапе оценки субградиента используют взвешенную сумму с ВФ $\Psi(i) = \frac{1}{\alpha x}$, регуляризованной по лифтинговой схеме [8], с начальной точкой, определенной на предыдущем этапе. Если координата минимума отличается от результата предыдущего этапа не более чем на δ, поиск заканчивается, иначе - масштаб α увеличивается на 1.

Таким образом, от оптимизации с помощью ВФ Хаара, обеспечивающей помехоустойчивость, переходят к оптимизации с помощью дифференциатора, обеспечивающего высокую точность (если $\alpha \to \infty$, то $\frac{1}{\alpha x}$ стремится к дифференциатору). После этого пересчитывают $\mu_{k,i}$ по (1), пока не выполнится условие останова.

Оценка чувствительности к локальным экстремумам и стартовой точке на этапе поиска оптимума проводилась с помощью функции Швефеля (с ложным глобальным минимумом). Точка старта выбиралась случайным образом. Методом градиентного спуска отыскивали ближайший к стартовой точке минимум, а разработанный метод кластеризации позволил достичь глобального минимума в 12 случаях из 15. Это доказывает повышение вероятности выхода в область глобального экстремума.

Помехоустойчивость в отличие от метода градиентного спуска исследовалась для классификации, проведенной после кластеризации в два этапа: в режиме обучения и в рабочем режиме.

Для двух классов в условиях помех в рабочем режиме оценена зависимость изменения суммарной вероятности ошибок первого и второго рода при увеличении относительной величины среднеквадратического отклонения (СКО) $q = \frac{q_{\partial}}{q_0 \cdot D}$., где q_{∂} – СКО рабочего режима, q_0 – СКО режима обучения (помеха распределена по нормальному закону с нулевым средним), D – расстояние между центрами кластеров обучающей выборки. В результате этого при классификации средний риск уменьшился от 3 до 30 раз при изменении q от 0,04 до 0,23 (для длины носителя ВФ $L = 14$) [5].

Полученные результаты позволяют рекомендовать разработанный метод кластеризации к применению в широком круге практически важных задач классификации и кластеризации при техническом диагностировании в случае многомерных параметров, высоком уровне помех и малых объемов выборок.

Список использованной литературы

1. Вятченин Д.А. Нечеткие методы автоматической классификации / Д. А. Вятченин. – Минск: УП «Технопринт», 2004. – 219 с.

2. Мандель И. Д. Кластерный анализ / И. Д. Мандель. – М.: Финансы и статистика, 1988. – 176 с.

3. Крылов В. Н. Иерархический субградиентный итеративный метод оптимизации в пространстве вейвлет- преобразования / В. Н. Крылов, Г. Ю. Щербакова // Электроника и связь – К.: – 2008. –№ 6. – С.28–31.

4. Поляк Б. Т. Введение в оптимизацию / Б. Т. Поляк. – М.: Наука, 1983. – 384 с.

5. Щербакова Г. Ю. Субградиентный метод классификации в пространстве вейвлет-преобразования для технической диагностики / Г.Ю.Щербакова // Електротехнічні та комп'ютерні системи. – 2010. – №1 (77). – С.136–142.

6. Babu G. P. Clustering with evolution strategies / G. P. Babu, M. N. Murti // Pattern recognition. – 1994.– V. 27. – № 2. – P. 321 – 329.

7. Bezdek J.C. Pattern recognition with fuzzy objective function algorithms / J.C. Bezdek. – NY: Plenum Press.

8. Krylov V.N. Contour images segmentation in space of wavelet transform with the use of lifting / V.N. Krylov, M. V. Polyakova // Op-

tical-electronic informatively-power technologies. – 2007. – № 2 (12). – P. 48 – 58.

9. Li J. M. Agglomerative fuzzy k-means clustering algorithm with selection of number of clusters [Electronic Resources]/ J. M. Li, M. K. Ng, Y. M. Cheung, J. Z. Huang // IEEE trans. on knowledge and data engineering. – 2008. – V.20. – № 11. – P. 1519 – 1534. – Режим доступа: http://www.comp.hkbu.edu.hk/~ymc/papers/journal /tkde08_publication_version.pdf

10. Yang M. S. A Gaussian kernel-based fuzzy c-means algorithm with a spatial bias correction [Electronic Resources] / M. S. Yang, H. S. Tsai // Pattern recognition letters. – 2008.– № 29. – P. 1713 – 1725. – Режим доступа : http://www.elsevier.com/locate/patrec

11. Yang M. S. A survey of fuzzy clustering [Electronic Resources] / M. S. Yang // Math. Comput. Modelling. – 1993. – V. 18. – № 11. – P. 1 – 16. – Режим доступа : http://www2.math.cycu.edu.tw.

Получено 20.10.2011.

References

1. Vyatchenyn D.A. Fuzzy methods for automatic classification. – Minsk: UP "Tehnoprynt", 2004. – 219 p.[in Russian].

2. Mandel I.D. Claster analysis. – Moscow: Finance and Statistics, 1988. – 176 p. [in Russian].

3. Krylov V.N., G.Yu.Shcherbakova. Subhradiyentnyy iterative optimization method in wavelet space-trans-ation / Scientific Papers of the Military Institute of the Kiev. Proc. Univ them. Shevchenko. – Kiev: 2008. – № 6. – P.28–31 [in Russian].

4. Polyak B.T. Introduction to Optimization. – Moscow: Nauka, 1983. – 384 [in Russian].

5. Shcherbakov G. Yu. Subgradient method of classification in the space of wavelet transform for technical diagnostics / Elektrotehnichni that komp'yuterni system. – 2010. – № 1 (77). – P. 136–142 [in Russian].

6. Babu G.P., Murti M.N. Clustering with evolution strategies / Pattern recognition. – 1994.– V. 27. – № 2. – P. 321 – 329 [in English].

7. Bezdek J.C. Pattern recognition with fuzzy objective function algorithms. – NY : Plenum Press [in English].

8. Krylov V.N., Polyakova M.V. Contour images segmentation in space of wavelet transform with the use of lifting / Optical-electronic informatively-power technologies. – 2007. – № 2 (12). – P. 48 – 58 [in English].

9. Li J. M., Ng M.K., Cheung Y. M., Huang J.Z. Agglomerative fuzzy k-means clustering algorithm with selection of number of clusters [Electronic Resources] / IEEE trans. on knowledge and data engineering. – 2008. – V.20. – № 11. – P. 1519 – 1534. – Режим доступа:http://www.comp.hkbu.edu.hk/~ymc/papers/journal /tkde08_publication_version.pdf.[in English].

10. Yang M. S., Tsai H.S. A Gaussian kernel-based fuzzy c-means algorithm with a spatial bias correction [Electronic Resources] / Pattern recognition letters. – 2008.– № 29. – P. 1713 – 1725. – Режим доступа : http://www.elsevier.com/locate/patrec [in English].

11. Yang M.S. A survey of fuzzy clustering [Electronic Resources] // Math. Comput. Modelling. – 1993. – V. 18. – № 11. – P. 1 – 16. – Режим доступа : http://www2.math.cycu.edu.tw [in English].

Щербакова Галина Юрьевна, к.т.н. каф. ЭСИКТ Одесск. нац. политехн. ун-та, тел.734-8621
e-mail: Galina_onpu@mail.ru

Крылов Виктор Николаевич, д-р техн. наук каф. ПМ и ИТБ Одесск. нац. политехн. ун-та, тел. 779-7453
e-mail:Viktor_Krylov@inbox.ru

Логвинов Олег Викторович, к.т.н., каф. ЭСИКТ Одесск. нац. политехн. ун-та, тел.734-8621