УДК 004.852

©2012. **К. К. Кадомский**

ПОВЫШЕНИЕ ЭФФЕКТИВНОСТИ ИНКРЕМЕНТНОЙ КЛАСТЕРИЗАЦИИ НЕЧЕТКИХ ДАННЫХ

Рассмотрена задача кластеризации данных динамических измерений. Эта задача решается статистическим инкрементным методом. Предложен последовательный инкрементный алгоритм кластеризации нечетких данных, в котором модель кластера и модель входного образа учитывают их центр и форму. Для оценки расстояния между моделями предложена модификация расстояния Махаланобиса, которая сохраняет эвклидово расстояние в случае одноточечных моделей и позволяет сократить вычисления по сравнению с использованием расстояния Баттачария. Предложенный алгоритм позволяет повысить эффективность кластеризации по сравнению с существующими инкрементными алгоритмами и повысить скорость кластеризации по сравнению с итеративным EM алгоритмом.

Ключевые слова: *инкрементная кластеризация, статистические модели данных, нечеткие данные, расстояние Махланобиса.*

1. Введение. Методы кластерного анализа используются для автоматической группировки данных и являются основой методов извлечения знаний из набора данных. В интеллектуальных системах управления [1] и поддержки принятия решений [2] возникает задача кластеризации данных динамических измерений, которая имеет следующие особенности. Во-первых, обработка данных должна производиться в режиме on-line, т.е. элементы обучающей выборки подаются по одному, и в каждый момент времени доступна лишь часть обучающей выборки. Во-вторых, количество кластеров не может быть оценено заранее. В-третьих, предъявляются жесткие требования к временной сложности алгоритма кластеризации.

В подобных задачах итеративные методы кластеризации, такие как EM алгоритм [3], не применимы. Существующие инкрементные алгоритмы, специально разработанные для задач динамического анализа данных, либо не адаптированы к обработке нечетких входных данных [4, 5, 6, 7], либо не учитывают форму кластеров [8, 9]. Кроме того, общим их недостатком является проблема соотношения стабильности – пластичности обучения [10].

Здесь предложен последовательный инкрементный алгоритм кластеризации нечетких данных, в котором модель кластера и модель входного образа учитывают центр и форму нечеткого множества. Этот алгоритм позволяет преодолеть указанные выше недостатки и повысить скорость кластеризации по сравнению с EM алгоритмом.

2. Формальная постановка задачи. Имеется конечный набор нечетких входных образов $\underset{\sim}{x}_l = \{x|\mu_{X_l}(x)\}$ из пространства входных образов P, заданных на базовом множестве X – пространстве четких входных образов. Каждый элемент x пространства X есть числовой вектор. Необходимо построить нечеткое разбиение множества входных образов на подмножества (кластеры) по принципу сходства в

смысле некоторой выбранной оценки расстояния между образами $d : P^2 \to [0; \infty)$. Входные образы предъявляются по одному, и для каждого нового образа $\underset{\sim}{x_l}$ необходимо:

а) построить нечеткое разбиение X на кластеры по данным $\{\underset{\sim}{x_1}, \underset{\sim}{x_2}, ..., \underset{\sim}{x_l}\}$,

б) определить степень принадлежности $\underset{\sim}{x_l}$ каждому из кластеров.

В статистической интерпретации данной задачи каждому обычному, либо нечеткому подмножеству – входному образу, либо кластеру – соответствует некоторое статистическое распределение C значений признаков. Набор входных образов рассматривается как последовательность независимых наблюдений $x_l = (x_{l1}, ..., x_{ln})^T$, $l = \overline{1, N_I}$ многомерной случайной величины $\mathrm{X} = (\mathrm{X}_1, ..., \mathrm{X}_n)^T$, где n – количество признаков в описании каждого входного образа, N_I – количество входных образов. Каждому кластеру C_k ставится в соответствие статистическая модель $\theta(\mathrm{C}_k)$ распределения элементов этого кластера в пространстве X. Эта модель может быть либо строгой статистической [11], либо нечеткой [8, 9], имеющей статистическую интерпретацию. Степень принадлежности образа x кластеру тогда есть вероятность $p(x|\theta(\mathrm{C}_k))$, а распределение X является суммой неизвестных распределений C_k. Задача кластеризации при этом сводится к оценке неизвестных параметров модели классификатора $\Theta = (M, \theta(\mathrm{C}_1), ..., \theta(\mathrm{C}_M))$ на основе наблюдений случайной величины X.

3. Статистические инкрементные методы кластеризации. Инкрементные алгоритмы [4, 6, 7, 10, 12] рассматривают каждый входной образ x_l независимо, используя его для модификации текущей модели классификатора $\Theta = (M, \theta(\mathrm{C}_1), ..., \theta(\mathrm{C}_M))$ согласно инкрементным соотношениям вида $\Theta' = f(\Theta, x_l)$. Инкрементный алгоритм жесткой кластеризации известен как алгоритм ведущего кластера (Sequential Leader Clustering, SLC) [12] и широко применяется в задачах сжатия обучающей выборки [5], обучения ИНС [13, 9] и нейроподобных сетей [7, 10]. В качестве модели кластера используется пара $\theta(\mathrm{C}_k) = \langle w_k, m_k \rangle$, где w_k – мощность кластера, и $m_k = \mathrm{E}(\mathrm{C}_k)$ – центроид. Каждый входной образ x_l либо относится к одному из существующих кластеров C_k, либо служит прототипом нового кластера: $\theta(\mathrm{C}_{new}) = \langle 1, x_l \rangle$. В первом случае параметры модели $\theta(\mathrm{C}_k)$ изменяются согласно инкрементным соотношениям

$$w'_k = w_k + 1,$$

$$m'_k = m_k + \eta(w_k)(x_l - m_k),$$

где $\eta(w_k)$ – функция скорости обучения, зависящая от количества элементов в кластере w_k и удовлетворяющая критериям статистической аппроксимации Дворецкого [14].

Выбор функции $\eta(w_k)$ рассматривается в работах [9, 15]. В работах [4, 9] алгоритм SLC обобщен на случай нечеткой кластеризации. Существуют также инкрементные алгоритмы (например, GenIc [4]), которые разбивают входной поток данных на окна, решая задачу кластеризации EM алгоритмом в пределах каждого окна отдельно, и затем последовательно объединяя результаты.

Инкрементные алгоритмы, как правило, не являются итеративными (обрабатывают входные образы либо окна однократно) и не требуют хранения в памяти всей обучающей выборки. Поэтому они имеют меньшую временную и емкостную сложность, по сравнению с EM алгоритмом. Однако существующие инкрементные алгоритмы имеют ряд недостатков. Так, алгоритмы, предложенные в работах [4, 5, 6, 7, 10], не адаптированы к обработке нечетких входных данных, а алгоритмы [8, 9] не поддерживают сложные нечеткие и статистические модели для представления кластеров данных, что не позволяет учесть форму кластеров. Также известна проблема соотношения стабильности – пластичности обучения при использовании инкрементных алгоритмов [7d].

Качество кластеризации существенно зависит от выбора типа модели и оценки расстояния между моделями. В качестве модели распределения $\theta(\mathrm{C})$ наиболее часто используется пара $\langle \mathrm{E(C)}, \mathrm{cov(C)} \rangle$, где E(C) – матожидание (центроид), cov(C) – ковариация [11]. Реже вместо матожидания и ковариации используются мода и вариация относительно моды [16], что позволяет уменьшить чувствительность алгоритма к случайным выбросам.

В простейшем случае расстояние между моделью C и точкой x есть расстояние (эвклидово, манхэттенское, либо расстояние Чебышева) между точками E(C) и x. Для учета размера и формы кластеров используются нормализованное эвклидово расстояние (1) и расстояние Махаланобиса (Mahalanobis) [17, 18] (2)

$$d^2_{EN}(\mathrm{C}_k,x) = \mathrm{diag}(\mathrm{cov}(\mathrm{C}_k))^{-1} \left(x - \mathrm{E}(\mathrm{C}_k)\right)^2, \tag{1}$$

$$d^2_{M}(\mathrm{C},x) = \left(x - \mathrm{E(C)}\right)^T \mathrm{cov}^{-1}(\mathrm{C}) \left(\mathrm{x} - \mathrm{E(C)}\right). \tag{2}$$

Для оценки расстояния между двумя моделями C_1 и C_2 используется расстояние Баттачария (Bhattacharyya) [19] (3) и производное от него расстояние Хеллингера (Hellinger) [18].

$$d_B(\mathrm{C}_1, \mathrm{C}_2) = -\ln \int \sqrt{p_1(x) p_2(x)} dx, \tag{3}$$

где $p_1(x)$ и $p_2(x)$ – плотность распределения C_1 и C_2, соответственно.

Если C_1 и C_2 имеют нормальное распределение, то несмещенная оценка расстояния Баттачария [19] между ними может быть получена как

$$d_B(\mathrm{C}_1, \mathrm{C}_2) = \frac{1}{8}(m_1 - m_2)^T S^{-1} (m_1 - m_2) - \frac{1}{2} \ln \left(\frac{\det S}{\sqrt{\det S_1 \det S_2}} \right), \tag{4}$$

где $m_1 = \mathrm{E}(\mathrm{C}_1), m_2 = \mathrm{E}(\mathrm{C}_2)$; $S = (\mathrm{cov}(\mathrm{C}_1) + \mathrm{cov}(\mathrm{C}_2))/2$.

Эти оценки требуют либо интегрирования по всему признаковому пространству (3), либо обращения матрицы ковариации для каждой пары (входной образ – кластер) (4). Кроме того, оценки Махаланобиса и Баттачария не определены в случае $\det \mathrm{cov(C)} \to 0$, например, для одноточечных распределений. Здесь предложена оценка расстояния между моделями, которая сохраняет эвклидово расстояние между точками и требует однократного обращения матрицы ковариации после каждого изменения модели кластера.

Также предложен последовательный инкрементный алгоритм кластеризации нечетких данных, который представляет кластеры и нечеткие входные образы эллипсоидами, произвольно ориентированными в пространстве X, и позволяет преодолеть указанные выше недостатки существующих инкрементных алгоритмов.

4. Модель данных. Каждому нечеткому входному образу и каждому нечеткому кластеру ставится в соответствие статистическое распределение C элементов данного нечеткого множества.

Если для функции принадлежности исходного множества выполняется условие нормировки

$$\int_{x\in X} \mu(x)dx \ = 1,$$

то плотность распределения C совпадает с функцией принадлежности μ. Если это условие не выполняется, то плотность распределения C есть $\mu(x)/\int \mu(x)dx$.

Распределение C задается параметрической моделью

$$\theta(\mathrm{C}) = \langle w(\mathrm{C}), \mathrm{E}(\mathrm{C}), \mathrm{cov}(\mathrm{C})\rangle \,,$$

где w(C) – количество наблюдений элементов распределения C (для входных образов эта величина равна 1); E(C) – оценка ожидания распределения C; cov(C) – оценка его матрицы ковариации [11].

Параметр E(C) задает центральный вектор распределения и центр исходного нечеткого множества. Параметр cov(C) определяет форму и размеры распределения, а также форму и размеры сечений уровня α [20] исходного нечеткого множества для каждого $\alpha \in (0;1)$. Каждое такое сечение является эллипсоидом, а длина и направления его осей определяются собственными числами и главными направлениями матрицы cov(C). В дальнейшем распределение C и его модель θ(C) считаются синонимами.

5. Расстояние в пространстве моделей. Здесь вводится модификация расстояния Махаланобиса между моделью распределения и точкой, которая сохраняет обычное эвклидово расстояние в случае одноточечного распределения. Затем эта оценка обобщается на случай расстояния между двумя различными моделями.

Рассмотрим модель распределения $\mathrm{C} = \langle w, m, S\rangle$, где $m = \mathrm{E}(\mathrm{C})$, $\mathrm{S} = \mathrm{cov}(\mathrm{C})$. Пользуясь свойствами ковариационной матрицы, нетрудно показать, что расстояние Махаланобиса между моделью C и точкой x (2) путем линейного преобразования сводится к нормированному эвклидову расстоянию:

$$d_M^2(\mathrm{C},x) = d_{EN}^2\left(U^T\mathrm{C},U^Tx\right) = \left(\Sigma^{-1}U^T(x-m)\right)^2,$$

где $U = (u_1, u_2, ..., u_n)$ – модальная матрица линейного оператора S; $u_1, u_2, ..., u_n$ – его собственные векторы; $\Sigma^2 = \mathrm{diag}(\sigma_1^2, ..., \sigma_n^2)$ – каноническая форма S.

Для сохранения эвклидова расстояния между одноточечными распределениями предлагается использовать обобщенное эвклидово расстояние

$$d_{EN*}(\mathrm{C},x) = \sqrt{f(\Sigma)(x-m)^2},$$

где $f(\sigma)$ – функция, монотонно убывающая на интервале $[0;\infty)$, такая, что $f(0)=1$, и $f(\sigma)\to 0$ при $\sigma\to\infty$.

Определение 1. Обобщенным расстоянием Махаланобиса между распределением $\mathrm{C}=\langle w,m,S\rangle$ и точкой x будем называть расстояние

$$d_{M*}(\mathrm{C},x)=d_{EN*}\left(U^T\mathrm{C},U^Tx\right). \tag{5}$$

В качестве $f(\sigma)$ предлагается использовать функцию вида

$$f(\sigma)=\alpha^2/(\alpha^2+\sigma^2),\alpha>0. \tag{6}$$

Тогда получим модификацию расстояния Махаланобиса

$$d^2_{M*}(\mathrm{C},x)=(x-m)^T(\mathrm{I}+\alpha^{-2}S)^{-1}(x-m), \tag{7}$$

которая сохраняет эвклидово расстояние в случае одноточечного распределения, и асимптотически приближается к обычному расстоянию Махаланобиса при $\sigma_i\to\infty$.

Вычисление расстояния по формуле (7) с использованием (6) более эффективно, чем в общем случае по определению (5), поскольку нахождение собственных векторов является более дорогостоящей операцией, по сравнению с обращением матрицы. В качестве $f(\sigma)$ можно использовать и другие убывающие функции, например $f(\sigma)=a/(a+\sigma)$, однако, в большинстве случаев вычислительная сложность при этом увеличивается. Так, в случае $f(\sigma)=a/(a+\sigma)$ получим

$$d^2_{M*}(\mathrm{C},x)=(x-m)^T(\mathrm{I}+\alpha^{-2}S+2\alpha^{-1}S^{1/2})^{-1}(x-m),$$

что за счет дорогостоящей операции нахождения квадратного корня матрицы имеет большую вычислительную сложность, чем (7).

Определение 2. Обобщенным расстоянием Махаланобиса между двумя распределениями $\mathrm{C}_1=\langle w_1,m_1,S_1\rangle$ и $\mathrm{C}_2=\langle w_2,m_2,S_2\rangle$ будем называть расстояние

$$d_{M*}(\mathrm{C}_1,\mathrm{C}_2)=2\frac{d_{M*}(\mathrm{C}_1,m_2)\cdot d_{M*}(\mathrm{C}_2,m_1)}{d_{M*}(\mathrm{C}_1,m_2)+d_{M*}(\mathrm{C}_2,m_1)}. \tag{8}$$

Поскольку, матрица $S'=\mathrm{I}+\alpha^{-2}S$ в (7) симметрична и неотрицательно определена, ее можно рассматривать как ковариацию некоторого распределения C'. Следовательно, $d_{M*}(\mathrm{C},x)=d_M(\mathrm{C}',x)=|x-m|/|r'|$, где r' – радиус эллипсоида, заданного уравнением $d_M(\mathrm{C}',x)=1$, по направлению $(x-m)$. Тогда расстояние (8) можно записать как

$$d_{M*}(\mathrm{C}_1,\mathrm{C}_2)=2\frac{|m_1-m_2|}{|r'_1|+|r'_2|}.$$

Таким образом, предложенное расстояние между двумя моделями имеет простую геометрическую интерпретацию (рис. 1).

6. Оценка параметров модели. Рассмотрим метод оценки параметров модели кластера $\theta(\mathrm{C})=\langle w,m,S\rangle$ на основе последовательных динамических наблюдений

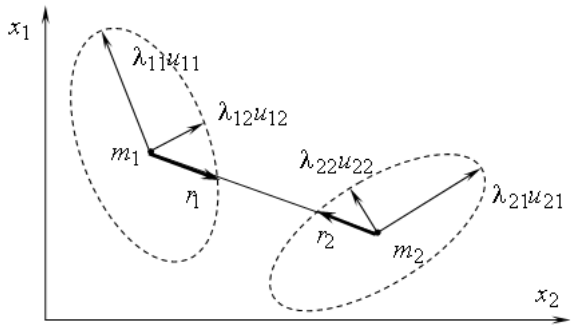

Рис. 1. Геометрическая интерпретация расстояния между двумя статистическими моделями

элементов этого кластера $\{x_r|\mu(x_r)\}_{r=1}^N$, где $x_r = (x_{r1}, x_{r2}, ..., x_{rn})$ – r-е наблюдение; $\mu(x_r)$ – степень уверенности, в том, что $x_r \in$ C. Наблюдения обрабатываются по одному, и на каждом шаге производится уточнение оценки параметров, полученной на предыдущем шаге с учетом нового наблюдения. В отличие от итеративных методов, таких как EM алгоритм, данный метод не требует выбора начальных значений параметров модели, не требует дополнительной памяти и многократной обработки обучающей выборки.

Поскольку параметр w есть мощность множества наблюдений, отнесенных к кластеру, то

$$w_N = \sum_{r=1}^{N} \mu(x_r).$$

Для каждого $k = \overline{1, N}$ плотность распределения C на выборке $\{x_r|\mu(x_r)\}_{r=1}^k$ есть

$$p_{\mathrm{C}}(x) = \mu(x) / \sum_{r=1}^{k} \mu(x_r), x \in \{x_1, ..., x_k\}.$$

Обозначим $p_r = p_{\mathrm{C}}(x_r)$. Выборочные оценки параметров m и S есть соответственно,

$$m_k = \sum_{r=1}^{k} x_r p_r, \tag{9}$$

$$S_{kij} = \sum_{r=1}^{k} (x_{ri} - m_{ki})(x_{rj} - m_{kj}) p_r, i, j = \overline{1, N_f}. \tag{10}$$

Обозначим $m = m_{k-1}$ и $S = S_{k-1}$; $m' = m_k$ и $S' = S_k$. Применяя к (9) условие нормировки $\sum_{r=1}^{k} p_r = 1$, для оценки m получим

$$m' = m \sum_{r=1}^{k-1} p_r + x_k p_k = (1 - p_k)m + p_k x_k. \tag{11}$$

Для оценки S путем преобразований (10) получим

$$\begin{aligned} S'_{ij} &= (1-p_k)S_{ij} + p_k(1-p_k)(x_{ki}-m_i)(x_{kj}-m_j) = \\ &= \left(1-\tfrac{\mu(x_k)}{w_k}\right)\left(S_{ij} + \tfrac{\mu(x_k)}{w_k}(x_{ki}-m_i)(x_{kj}-m_j)\right). \end{aligned} \tag{12}$$

7. Алгоритм кластеризации. Данный алгоритм использует расстояние (8) для оценки степени принадлежности входного образа кластеру и инкрементные соотношения (11-12) для выборочной оценки параметров модели кластера. Алгоритм кластеризации имеет следующий вид.

1. Инициализировать модель первого кластера параметрами первого входного образа $\Theta = (1, \theta(\underset{\sim}{x}_1))$.

2. При получении нового входного образа $\underset{\sim}{x}$ преобразовать его к статистической модели $\theta(\underset{\sim}{x}) = \langle w_x, m_x, S_x \rangle$.

3. Для каждого $k = \overline{1, M}$ вычислить расстояние $d_k = d_{M*}(\mathrm{C}_k, \underset{\sim}{x})$ (8).

4. Если все вычисленные расстояния превышают заданный порог $d_{\max}$, то добавить новый кластер, инициализировать его параметрами $\langle w_x, m_x, S_x \rangle$ и перейти к следующему входному образу.

5. В противном случае вычислить степени принадлежности входного образа каждому из кластеров

$$\mu_{\mathrm{C}_k}(\underset{\sim}{x}) = d_k^{-p} / \sum_{j=1}^{M} d_j^{-p},$$

где $p > 1$ – параметр, определяющий степень нечеткости алгоритма кластеризации. При малых p имеем нечеткую кластеризацию, а в предельном случае $p \to \infty$ – жесткую.

6. Составить подмножество активных кластеров, для которых $\mu_{\mathrm{C}_k}(\underset{\sim}{x}) \geq \mu_{\min}$. Для каждого активного кластера уточнить оценки параметров модели по формулам:

$$w' = w + \mu(x_k),$$

$$\beta = \max\{\mu(x_k)/w', \beta_{\min}\mu(x_k)\},$$

$$m' = (1-\beta)m + \beta x_k,$$

$$S'_{ij} = (1-\beta)\left(S_{ij} + \beta(x_{ki}-m_i)(x_{kj}-m_j)\right),$$

где $\beta_{\min} \in [0; 1)$ – параметр, определяющий скорость забывания предыдущих наблюдений.

7. Вычислить обратную матрицу S^{-1}.

8. Результаты. Предложенный алгоритм кластеризации апробирован в классификаторе нечетких описаний ситуации в составе системы управления мобильным роботом Lego® Mindstorms® NXT 2.0. На рис. 2 приведены две проекции уменьшенной обучающей выборки, состоящей из 80 3-мерных образов, и результаты нечеткой кластеризации предложенным алгоритмом с параметрами ($p = 2, d_{\max} = 0.2, \beta_{\min} = 0.02, \mu_{\min} = 0.0025$).

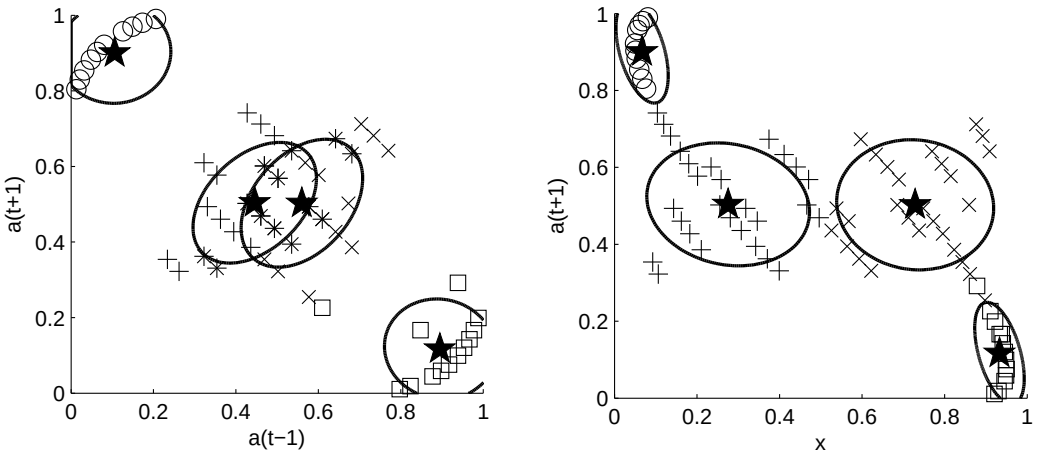

Рис. 2. Результаты кластеризации инкрементным алгоритмом

В таблице 1 даны экспериментальные показатели предложенного алгоритма в сравнении с итеративными алгоритмами k-means и fuzzy c-means, а также инкрементным алгоритмом SLC. В качестве критерия качества кластеризации использовался индекс Дэвиса-Боулдина (Davies–Bouldin index) [21].

Таблица 1. Сравнение методов кластеризации.

Алгоритм	**SLC**	**k-means**	**fuzzy c-means**	**жесткий инкр.**	**нечеткий инкр.**
Максимальное количество итераций	1	11	18	1	1
Индекс Дэвиса-Боулдина	0.48	0.31	0.37	0.31	0.38

Предложенный алгоритм обеспечивает более высокое качество кластеризации (меньшее значение индекса Дэвиса-Боулдина) по сравнению с алгоритмом SLC. В предложенном алгоритме входные данные обрабатываются однократно, что обеспечивает более высокую скорость работы, по сравнению с алгоритмами k-means и fuzzy c-means. Кроме того, предложенный алгоритм обеспечивает динамическую оценку количества кластеров.

9. Выводы. Рассмотрена задача повышения эффективности кластеризации дан-

ных нечетких динамических измерений. Выделены основные недостатки существующих инкрементных методов кластеризации [4, 5, 6, 7, 8, 9, 10]. Использована статистическая модель данных, основанная на ожидании и ковариации. Эта модель представляет кластеры и нечеткие входные образы эллипсоидами, произвольно ориентированными в пространстве четких входных образов. Предложен метод оценки параметров модели кластера на основе последовательных динамических наблюдений элементов этого кластера. Предложены оценки расстояния между моделью и точкой, а также между двумя различными моделями, которые сводятся к евклидовому расстоянию в случае одноточечных моделей. На основе этих моделей и методов предложен последовательный инкрементный алгоритм кластеризации нечетких данных, который позволяет повысить качество кластеризации по сравнению с существующим инкрементным алгоритмом SLC, а также повысить скорость обработки данных по сравнению с итеративными алгоритмами k-means и fuzzy c-means.

1. *Каргин А.А.* Введение в интеллектуальные машины. Книга 1. Интеллектуальные регуляторы. – Донецк: Норд-Пресс, ДонНУ, 2010. -– 526 с.
2. *Marakas G.M.* Decision support systems in 21st century. – US edition. – Upper Saddle River, N.J. : Prentice Hall, 1999. – 528 p.
3. *Gupta M.R., Chen Y.* Theory and use of the EM algorithm // Foundations and trends in signal processing. – 2010. – Vol. 4. No. 3. – PP. 223–296.
4. *Gupta C., Grossman R.* Genic: a single pass generalized incremental algorithm for clustering // Proceedings of the Fourth SIAM International Conference on Data Mining. – 2004. – PP. 147-153.
5. *Lin J., Vlachos M., Keogh E., Gunopulos D.* Iterative incremental clustering of time series // Advances in database technology. – 2004. – Vol. 2992. – PP. 521-522.
6. *Li D., Simske S.* Training set compression by incremental clustering // Journal of pattern recognition research. – 2011. – Vol 6. No 1. – PP. 56-64.
7. *Anagnostopoulos G.C., Georgiopoulos M.* Ellipsoid ART and ARTMAP for incremental clustering and classification // Neural networks. – 2001. – Vol. 2. – PP. 1221-1226.
8. *Nefti S. A, Oussalah M., Rezgui Y.* A modified fuzzy clustering for documents retrieval: application to document categorization // Journal of the Operational Research Society. – 2009. – Vol. 60. No. 3. – PP. 384-394.
9. *Бодянский Е.В., Волкова В.В., Махиборода В.В.* Нейронная сеть Т. Кохонена с нечетким выводом и алгоритм ее самообучения // Збірник наукових праць Харківського університету Повітряних Сил. – 2009. – Вип. 2 (**20**). – С. 74-78.
10. *Carpenter G. A., Grossberg S.* Adaptive resonance theory // The handbook of brain theory and neural networks / M. A. Arbib (ed.). – 2nd edition. – Cambridge, MA : MIT Press, 2003. – PP. 87-90.
11. *Wasserman L.* All of statistics: a concise course in statistical inference. – New York : Springer, 2004. – 561 p.
12. *Hartigan J.A.* Clustering Algorithms. – New York: Wiley, 1975. – 351 p.
13. *Haykin S.* Neural networks and learning machines. – 3rd edition. – Upper Saddle River, NJ : Prentice Hall, 2008. – 936 p.
14. *Dvoretzky A.* On stochastic approximation // 3-rd Berkeley symp. math. statistics and probability : proc. of. – 1956. – Vol. 1. – PP. 39-55.
15. *Goodwin G.C., Ramadge P.J., Caines P.E.* A globally convergent adaptive predictor // Automatica. – 1981. – Vol. 17. No. 1. – PP. 135-140.
16. *Li J., Ray S., Lindsay B.G.* A nonparametric statistical approach to clustering via mode identification // Journal of machine learning research. – 2007. – Vol. 8. – PP. 1687-1723.
17. *Maesschalck R. de., Jouan-Rimbaud D., Massart D.L.* The Mahalanobis distance // Chemometrics and intelligent laboratory systems. – 2000. – Vol. 50. Issue 1. – PP. 1–18.
18. *Vaart, van der A. W.* Asymptotic statistics. – Cambridge, UK : Cambridge University Press, 2000.

– 460 p.
19. *Nielsen F., Boltz S.* The Burbea-Rao and Bhattacharyya centroids // Information theory : IEEE trans. on. – 2011. – Vol. 57 (**8**). – PP. 5455-5466.
20. *Раскин Л.Г., Серая О.В.* Нечеткая математика. Основы теории. Приложения. – Х.: Парус, 2008. – 352 с.
21. *Romesburg H.C.* Cluster analysis for researchers. – New York : Lulu Press, 2004. – 344 p.

C. Kadomsky

Efficient incremental clustering of fuzzy data.

The problem of dynamic data clustering is addressed. This problem is solved by statistical incremental method. The sequential incremental fuzzy data clustering algorithm is proposed, in which the cluster model and the input model account for their center and shape. For estimating distance between models the modification of Mahalanobis distance is proposed, which preserves Euclidean distance in case of single-point models and allows reducing calculations in comparison with the use of Bhattacharyya distance. The proposed algorithm allows to improve clustering efficiency in comparison with existing incremental algorithms, and to improve clustering speed in comparison with iterative EM algorithm.

Keywords: *incremental clustering, statistical data models, fuzzy data, Mahalanobis distance.*

К. К. Кадомський

Підвищення ефективності інкрементної кластеризації нечітких даних.

Розглянуто задачу кластеризації даних динамічних вимірів. Ця задача вирішується статистичним інкрементним методом. Запропоновано послідовний інкрементний алгоритм кластеризації нечітких даних, в якому модель кластера та модель вхідного образу враховують їх центр і форму. Для оцінки відстані між моделями запропоновано модифікацію відстані Махаланобіса, яка зберігає евклідову відстань у випадку одноточкових моделей і дозволяє скоротити обчислення в порівнянні з використанням відстані Баттачарія. Запропонований алгоритм дозволяє підвищити ефективність кластеризації в порівнянні з існуючими інкрементними алгоритмами та підвищити швидкість кластеризації в порівнянні з ітеративним EM алгоритмом.

Ключові слова: *інкрементна кластеризація, статистичні моделі даних, нечіткі дані, відстань Махланобіса.*

Донецкий национальный ун-т
kadomsky@ukr.net

Получено 11.05.12