УДК 519.8

Э.Н. НАСИБОВ

ПОМЕХОУСТОЙЧИВЫЙ АЛГОРИТМ РЕШЕНИЯ ПРОБЛЕМЫ НЕЧЕТКОЙ КЛАСТЕРИЗАЦИИ НА БАЗЕ МЕТОДА НЕЧЕТКИХ СВЯЗАННЫХ ТОЧЕК

Ключевые слова: *кластеризация, нечеткие связанные точки (FJP), нечеткое связанное множество, помехоустойчивость.*

ВВЕДЕНИЕ

Задача кластеризации является одной из основных задач, используемых в современной технологии интеллектуальной обработки данных с применением компьютеров [1, 2]. Общая философия кластеризации базируется на разбиении исходного массива данных на гомогенные (схожие) классы относительно близости свойств. При этом элементы в одном классе являются максимально близкими, а элементы в различных классах являются максимально далекими по отношению один к другому.

В классической задаче кластеризации границы классов являются четкими и любой элемент может принадлежать только одному из имеющихся классов. Однако на практике границы между классами, как правило, не могут быть определены четко и некоторые элементы могут одновременно принадлежать различным классам с отличными от нуля степенями принадлежностей. В таких случаях адекватность задачи кластеризации, подобно задаче идентификации [3–5], может быть повышена с применением теории нечетких множеств [2]. В настоящей статье рассматривается другая концепция задачи кластеризации — с использованием так называемого метода нечетких связанных точек (Fuzzy Joint Points — FJP) [6]. Основное отличие такого метода заключается в том, что понятие размытости кластеризации рассматривается с иерархической точки зрения. А именно нечеткость кластеризации понимается в том смысле, насколько детально рассматриваются свойства элементов при формировании множества схожих элементов. Понятно, что чем более детально рассматриваются свойства, тем больше отличий выявляется между элементами. Если свойства рассматриваются более размыто, т.е. не вдаваясь в детали, то элементы имеют больше схожести. В таком случае размытость кластеризации определяется «детальностью» учета исследуемых свойств. Тогда при минимальной степени размытости все элементы в какой-то степени будут отличаться один от

другого и каждый элемент должен рассматриваться как отдельный класс, а при максимальной степени размытости все элементы будут похожи один на другого и, следовательно, все они будут принадлежить одному классу. При какой-то степени размытости между 0 и 1 некоторые элементы с близкими свойствами будут похожи один на другого и принадлежать одному и тому же классу, а более удаленные элементы будут принадлежать различным классам.

Основными вопросами в проблеме определения нечеткой кластеризации являются: а) нахождение оптимального количества гомогенных классов; б) формирование начального разбиения исходных элементов на классы; в) непосредственные алгоритмы кластеризации с итеративным улучшением разбиения.

Существует достаточное число публикаций по указанным вопросам [7–11]. Среди рассмотренных методов наиболее часто используются метод ближайщих K-соседей (K-Nearest Neighbors — KNN) и метод холмов (Mountain Method) [8–10]. Однако они имеют некоторые недостатки. Так, в методе KNN требуется заранее задавать количество кластеров, кроме того, разбиение является достаточно грубым. Основным недостатком метода холмов — сложность при его вычислении.

Рассмотренный в настоящей статье метод FJP свободен от вышеперечисленных недостатков [6]. Одним из преимуществ указанного метода — существование механизма автоматического нахождения количества классов, который отсутствует в большинстве других алгоритмов иерархической кластеризации. Этот метод может также распознавать отделенные элементы как отдельные классы, что свидетельствует в пользу метода FJP, поскольку удаленные от массы элементы могут быть не игнорированы как точки шума, а рассмотрены как самостоятельные индивидуальные классы. Это может иметь значение, например, при исследовании в области медицины, биологии и т.п., где аномальные случаи имеют определенный интерес. В настоящей статье исследуется условие корректного распознавания классов алгоритмом FJP и предлагается помехоустойчивый вариант этого алгоритма, в котором его чувствительность к шуму может быть отрегулирована.

1. ПОСТАНОВКА ЗАДАЧИ

Пусть имеется множество точек $X = \{x_1, x_2, \ldots, x_n\}$, где $x_i \in E^p$, $i = \overline{1, n}$. Требуется разделить множество X на гомогенные классы, т.е. произвести кластеризацию множества X. Другими словами, требуется вычислить подмножества $X^1, X^2, \ldots, X^k$ такие, что

$$\# i \neq j \quad X^i \cap X^j = \varnothing \tag{1}$$

и

$$\bigcup_{i=1}^{k} X^i = X. \tag{2}$$

Отметим, что подмножества $X^1, X^2, \ldots, X^k$ являются множествами в классическом смысле, т.е. без нечеткости.

Количество классов k заранее неизвестно. Поэтому требуется также определить количество классов, адекватно отражающее структуру данных. В рассматриваемом случае подходящее значение k будет определяться на основе оптимальности некоторой функции, базирующейся на максимальности межклассовых расстояний.

Как было сказано во введении, элементы внутри одного класса должны быть расположены максимально близко, а различные классы должны располагаться

как можно дальше один от другого. При этом близость элементов может определяться на основе различных метрик. Отметим, что большинство методов кластеризации на базе дистанции используют классическую эвклидову дистанцию, т.е. дистанция между обычными точками a и b p-мерного пространства E^p определяется как

$$d(a, b) = \sqrt{\sum_{i=1}^{p} (a_i - b_i)^2}. \tag{3}$$

Следует отметить, что существуют также методы, где используются другие более обобщенные дистанции. Например, в [12] рассматривается проблема кластеризации с применением алгоритма нечеткой кластеризации (Fuzzy c-Means — FCM) на базе масштабируемой по направлениям дистанции и демонстрируются некоторые преимущества такой дистанции. Однако в настоящей статье будем использовать дистанцию в смысле (3).

2. НЕЧЕТКИЕ СВЯЗАННЫЕ ТОЧКИ И ИХ СВОЙСТВА

Обозначим $F(E^p)$ множество всех нечетких подмножеств пространства E^p. Пусть $\mu_A : E^p \to [0,1]$ обозначает функцию принадлежности нечеткого множества $A \in F(E^p)$.

Сначала напомним некоторые понятия, введенные в работе [6].

Определение 1. Нечеткой точкой конусной формы $A = (a, R) \in F(E^p)$ пространства E^p будем называть нечеткое множество (рис. 1), функция принадлежности которого определяется как

$$\mu_A(x) = \begin{cases} 1 - \dfrac{d(x, a)}{R} & d(x, a) \le R, \\ 0 & \text{в противном случае} \end{cases} \tag{4}$$

Здесь $a \in E^p$ является центром нечеткой точки A, а $R \in E^1$ является радиусом его носителя $\operatorname{supp} A$, где $\operatorname{supp} A = \{x \in E^p \mid \mu_A(x) > 0\}$. Очевидно, что α-уровневое множество нечеткой точки конусной формы $A = (a, R)$ будет множество в классическом смысле

$$A_\alpha = \{x \in E^p \mid \mu_A(x) \ge \alpha\} = \{x \in E^p \mid d(x, a) \le R \cdot (1 - \alpha)\} \tag{5}$$

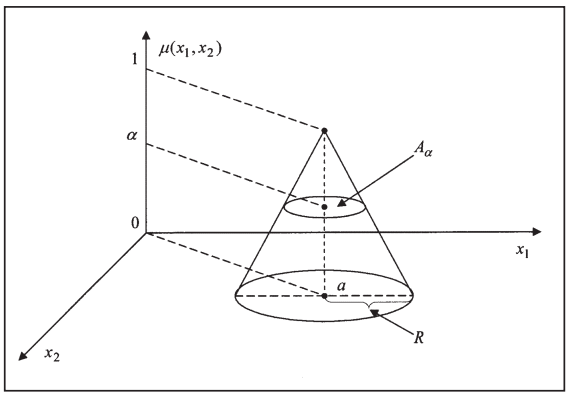

Рис. 1. Нечеткая точка конусной формы $A = (a, R) \in P(E^2)$ в пространстве E^2

Далее будем употреблять сокращенный термин «нечеткая точка», подразумевая при этом нечеткую точку конусной формы $(\bullet, R)$, определенную как (4).

Пусть $A = (a, R)$ и $B = (b, R)$ — нечеткие точки из множества $X \subset F(E^p)$. Определим нечеткое отношение $T : X \times X \to [0,1]$ между элементами множества X, степень выполнения которого определяется как

$$T(A, B) = \max\left\{1 - \frac{d(a, b)}{2R},\ 0\right\}, \tag{6}$$

где $a \in E^p$ и $b \in E^p$ являются центрами нечетких точек A и B соответственно (рис. 2). Из отношения (6) следует, что при $T(A, B) \in (0, 1]$ справедливо равенство

$$d(a, b) = 2R(1 - T(A, B)). \tag{7}$$

Понятно, что отношение T является рефлексивным, т.е. $\forall A \in X$ выполняется равенство $T(A, A) = 1$.

Определение 2. Пусть A и B являются нечеткими точками из множества $X \subset F(E^p)$. Если при фиксированном значении $\alpha \in (0, 1]$ удовлетворяется неравенство

$$T(A, B) \geq \alpha, \tag{8}$$

то точки A и B назовем α-соседними нечеткими точками: $A \sim_\alpha B$ (см. рис. 2).

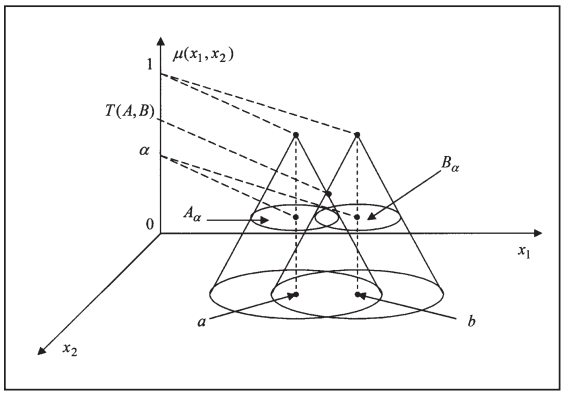

Рис. 2. Нечеткие α-соседние точкм $A = (a,R)$ и $B = (b,R)$ в пространстве E^2

Определение 3. Если при фиксированном значении $\alpha \in (0, 1]$ между нечеткими точками A и B имеется последовательность нечетких α-соседних точек $C^1, \ldots, C^k$, $k \geq 0$, т.е.

$$A \sim_\alpha C^1,\ C^1 \sim_\alpha C^2, \ldots,\ C^{k-1} \sim_\alpha C^k \text{ и } C^k \sim_\alpha B,$$

то нечеткие точки A и B назовем нечеткими α-связанными точками.

Определение 4. Пусть $X \subset F(E^p)$ является множеством нечетких точек. Если при фиксированном значении $\alpha \in (0, 1]$ для любых $A, B \in X$ нечеткие точки A и B являются α-связанными, то множество X назовем нечетким α-связанным множеством.

Лемма 1. Для того чтобы при фиксированном $\alpha \in (0, 1]$ нечеткие точки $A = (a, R)$ и $B = (b, R)$ были α-соседними, необходимо и достаточно выполнение неравенства

$$d(a, b) \leq 2R(1 - \alpha), \tag{9}$$

где $d(a, b)$ — расстояние между центрами нечетких точек A и B.

Лемма 2. Для того чтобы для некоторого $\alpha \in (0, 1]$ нечеткие точки A и B были α-соседними, необходимо и достаточно выполнение отношения

$$A_\alpha \cap B_\alpha \neq \varnothing. \tag{10}$$

Доказательства этих утверждений имеются в работе [6].

Пусть отношение $\hat{T}: X \times X \to [0, 1\ ]$ является транзитивным замыканием нечеткого отношения $T: X \times X \to [0, 1]$. При определении транзитивного замыкания будем предполагать классическую max–min-композицию нечетких отношений. Следующая теорема задает метод определения α-связанности между любыми заданными нечеткими точками.

Теорема 1. Небходимым и достаточным условием α-связанности любых

нечетких точек $A, B \in X$ конечного множества X при фиксированном $\alpha \in (0, 1]$ является неравенство

$$\hat{T}(A, B) \geq \alpha. \tag{11}$$

Доказательство теоремы имеется в работе [6].

3. ПОМЕХОУСТОЙЧИВЫЙ АЛГОРИТМ МЕТОДА НЕЧЕТКИХ СВЯЗАННЫХ ТОЧЕК

Для решения задачи (1), (2) предлагается следующий алгоритм, являющийся помехоустойчивым вариантом алгоритма нечетких связанных точек (Fuzzy Joint Points — FJP), изложенного в работе [6]. Этот алгоритм вычисляет подходящее значение четкости α, соответствующее количество классов k, а также разбиение исходного множества $X = \{x_1, x_2, \ldots, x_n\}$ на классы $X^1, X^2, \ldots, X^k$, удовлетворяющие отношениям (1) и (2). При этом $X^1, X^2, \ldots, X^k$ являются нечеткими α-связанными множествами.Пусть $N(x)$ — нечеткое соседство заданной точки $x \in X$ на основе отношения T, т.е.

$$N(x) = \{(y, T(x, y)) | \, y \in X\} \tag{12}$$

$$N(x, \varepsilon_1) = \{y \in X | \, T(x, y) \geq \varepsilon_1\} \tag{13}$$

является ε_1-уровневым множеством нечеткого множества $N(x)$, т.е. нечетким ε_1-соседством точки $x \in X$.

Определение 5. Точка $x \in X$ называется точкой шума с параметрами $\varepsilon_1, \varepsilon_2$ для заданных $\varepsilon_1 > 0$ и $\varepsilon_2 > 0$, если удовлетворяется неравенство $\text{card}\, N(x, \varepsilon_1) < \varepsilon_2$, где $\text{card}\, N(x, \varepsilon_1) = \sum\limits_{y \in N(x, \varepsilon_1)} T(x, y)$ является нечеткой мощностью множества $N(x, \varepsilon_1)$.

Предложенный ниже алгоритм FJP является помехоустойчивым. Если в данном алгоритме для заданной точки нечеткая мощность его нечеткого ε_1-соседства меньше ε_2, то эта точка воспринимается в качестве шума. Отметим, что настройкой параметров ε_1 и ε_2 можно регулировать чувствительность алгоритма FJP к помехам. Очевидно, что принимая $\varepsilon_2 = 0$, можно заглушить помехоустойчивость алгоритма.

Помехоустойчивый FJP-алгоритм

Шаг 1. Вычислить $d_{ij} := d(x_i, x_j)$, $i, j = \overline{1, n}$; $d_{\max} := \max\limits_{i, j = \overline{1, n}} d_{ij}$; $\varepsilon := 0{,}01 \cdot \min\limits_{i, j = \overline{1, n}} d_{ij}$. Установить значения ε_1 и ε_2. Пусть $\alpha_0 := 1$.

Шаг 2.Вычислить нечеткое отношение $T_{ij} := 1 - \dfrac{d_{ij}}{d_{\max}}$, $i, j = \overline{1, n}$.

Шаг 3.Вызвать процедуру *NoiseFilter*$(\varepsilon_1, \varepsilon_2)$ для разбиения исходного множества X на ядро X_{core} и шум X_{noise}, т.е. $X = X_{core} \cup X_{noise}$, $X_{core} \cap X_{noise} = \varnothing$. Вычислить транзитивное замыкание $\hat{T}$ отношения T на множестве X_{core}.

Шаг 4.Пусть n_c — количество элементов X_{core}; обозначить: $y_i := x_i$, $i = \overline{1, n_c}$; $t := 1$; $k := n_c$.

Шаг 5. Вычислить $d(y_i, y_j) = \min\{d(x', x'') | \, x' \in y_i, x'' \in y_j\}$, $i, j = \overline{1, k}$;

$$d_t := \min_{i \neq j} d(y_i, y_j); \; \alpha_t := \max\left\{1 - \frac{d_t + \varepsilon}{d_{\max}}, \; 0\right\}.$$

Шаг 6. Вызвать процедуру *Clusters*(α_t) для вычисления нечетких α_t-связанных множеств $X^1, X^2, \ldots, X^k$ с нечеткими точками конусной формы $\left(x_i, \dfrac{d_{\max}}{2}\right)$, $i = \overline{1, n_c}$, и для определения количества k этих множеств для заданного значения α_t.

Шаг 7. Если $k > 1$, то обозначить $y_i := X^i$, $i = \overline{1, k}$; $t = t + 1$ и перейти на шаг 5. Если $k = 1$, то перейти на шаг 8.

Шаг 8. Вычислить $\Delta\alpha_i := \alpha_i - \alpha_{i+1}$, $i = \overline{0, t-1}$; $z := \arg \max\limits_{i = 0, t-1} \Delta\alpha_i$; $\overline{\alpha} := \alpha_z - \varepsilon$.

Шаг 9. Вызвать процедуру *Clusters*($\overline{\alpha}$) с параметром $\overline{\alpha}$.

Шаг 10. $\overline{\alpha}$ — оптимальное значение уровня четкости кластеризации; $\overline{k}$ — оптимальное количество кластеров; $X^1, \ldots, X^{\overline{k}}$ — оптимальное разбиение множества X.

Шаг 11. Для каждого элемента $x \in X_{noise}$ повторить шаг 12.

Шаг 12. Вычислить $k^* = \arg\min\{dist(x, X^k) \mid k = 1, \ldots, \overline{k}\}$; элемент x отнести к множеству X^{k^*}.

End.

Процедура *NoiseFilter*($\varepsilon_1, \varepsilon_2$) используется для разбиения исходного множества X на два непересекающиеся множества ядра X_{core} и шума X_{noise}, т.е. $X = X_{core} \cup X_{noise}$, $X_{core} \cap X_{noise} = \varnothing$.

Процедура *NoiseFilter*($\varepsilon_1, \varepsilon_2$):

Входные параметры: ε_1 и ε_2.

Выходные параметры: множества X_{core} и X_{noise}.

Шаг 1. Пусть $X \equiv \{x_1, x_2, \ldots, x_n\}$ является множеством исходных точек, $X_{noise} = \varnothing$.

Шаг 2. Для каждой точки $x \in X$ повторить шаги 3 и 4.

Шаг 3. Вычислить $card\, N(x, \varepsilon_1) = \sum\limits_{y \in N(x, \varepsilon_1)} T(x, y)$.

Шаг 4. Если $cardN(x, \varepsilon_1) < \varepsilon_2$, то x пометить как точку шума: $X_{noise} := X_{noise} \cup \{x\}$.

Шаг 5. Обозначить $X_{core} := X \setminus X_{noise}$.

Шаг 6. Возвратить множества X_{core} и X_{noise}.

End.

Процедура *Clusters*(α) разбивает множество $X = \{x_1, x_2, \ldots, x_n\}$ на нечеткие α-связанные множества для входного параметра α, возвращает эти множества и количество этих множеств.

Процедура *Clusters*(α):

Входной параметр: α.

Выходные параметры: множества $X^1, X^2, \ldots, X^k$, отражающие гомогенные классы нечетких α-связанных точек; k — количество этих классов.

Шаг 1. $S := X = \{x_1, x_2, \ldots, x_n\}$; $k := 1$.

Шаг 2. Выбирать элемент $A \in S$ множества S. Образовать множества: $X^k := \{B \in S | \hat{T}(A, B) \geq \alpha\}$; $S := S \setminus {}^k$.

Шаг 3. Если $S \neq \varnothing$, то обозначить $k := k + 1$ и перейти к шагу 2; в противном случае перейти к шагу 4.

Шаг 4. Возвратить множества $X^1, X^2, \ldots, X^k$ и количество k этих множеств.

End.

Отметим, что в алгоритме FJP для нормализации нечеткого отношения $T : X \times X \to [0, 1]$ радиус рассматриваемых нечетких точек предполагается в виде

$$R = \frac{\max\{d(x_i, x_j) | x_i, x_j \in X\}}{2} \equiv \frac{d_{\max}}{2}.$$

Таким образом, $\# A, B \in X$ степень выполнения отношения T определится как

$$T(A, B) = 1 - \frac{d(a, b)}{d_{\max}}; \tag{14}$$

следовательно, $d(a, b) = d_{\max}(1 - T(A, B))$.

Пусть X^k, $k = \overline{1, t}$, — гомогенные классы, сформированные в результате кластеризации. Зададим следующие обозначения (см. рис. 3):

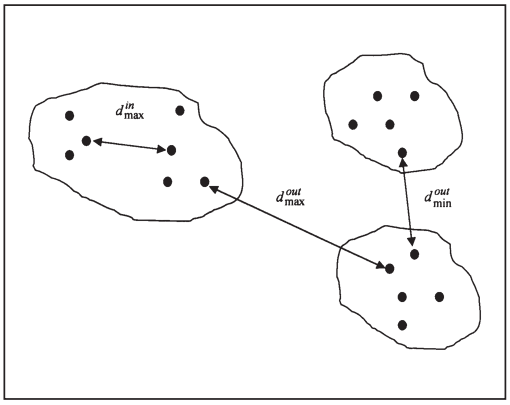

Рис. 3. Расположение гомогенных множеств, вызываемых алгоритмом FJP

$$d_k^{in} = d_{\max} \cdot (1 - \min_{x,\, y \in Xk} \hat{T}(x, y)),$$

$$d^{in}_{\max} = \max_k d_k^{in},$$

$$d^{out}_{\min} = \min_{i \neq j} d(X^i, X^j),$$

$$d^{out}_{\max} = \max_{i \neq j} d(X^i, X^j),$$

где $\hat{T}$ — транзитивное замыкание отношения T.

Критерий для нахождения оптимальной структуры классов основывается на максимальности широты изменения α-уровней, для которых количество классов не меняется. Другими словами, в качестве оптимальной принимается структура классов, для которой удовлетворяется

$$V_{FJP} \equiv d^{out}_{\min} - d^{in}_{\max} \to \max. \tag{15}$$

В табл. 1 также отражены другие существующие в литературе основные критерии определения оптимального количества классов при кластеризации [7]. При этом использованы следующие обозначения: u_{ij} — степень принадлежности элемента x_j к i-му классу, v_i — центр класса i, $\delta(u_i)$ — диаметр класса i, m_X — центр множества всех точек X.

Таблица 1. Критерии для определения оптимального количества классов

Название критерия	Функциональное выражение	Наилучшая кластеризация
Коэффициент разбиения (Partition coefficient)	$V_{PC} = \frac{1}{n}\sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}^2$	$\max(V_{PC}, U, c)$
Энтропия кластеризации (Classification entropy)	$V_{CE} = -\frac{1}{n}\sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}\log_a u_{ij}$	$\min(V_{CE}, U, c)$
Критерии Фукуяно–Сугено (Fukuyamo-Sugeno criteria)	$V_{FS_m} = \sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}^m [d^2(x_j, v_i) - d^2(m_X, v_i)]$	$\min(V_{FS}, U, c)$
Индекс разделения (Separation index)	$V_{SI} = \frac{\max\limits_i \min\limits_{k, k\neq i} d(u_i, u_k)}{\min\limits_i \delta(u_i)}$	$\max(V_{SI}, U)$
Критерий FJP (FJP criteria)	$V_{FJP} = d_{\min}^{out} - d_{\max}^{in}$	$\max(V_{FJP}, U)$

Приведенная ниже теорема определяет достаточное условие корректной кластеризации данных с помощью алгоритма FJP.

Теорема 2.Если имеется некоторая скрытая структура множества X с возможным разбиением (1), (2), удовлетворяющим соотнешениям

$$\frac{d_{\max}^{in}}{d_{\min}^{out}} < \frac{1}{2} < \frac{d_{\min}^{out} - d_{\max}^{in}}{d_{\max}^{out}}, \tag{16}$$

то алгоритм FJP распознает это разбиение как оптимальное.

Доказательство. Предположим, что имеется скрытая структура с возможным разбиением, удовлетворяющим соотношениям (16). Значения α, полученные в результате циклического применения шагов 5–7 помехоустойчивого алгоритма FJP, обозначим $\alpha_0, \alpha_1, \ldots, \alpha_t$. Согласно формуле (7) каждому значению степеней α_i, $i = \overline{0, t}$, соответствует некоторое значение радиуса окружности, которое определяет α-уровневое множество. По утверждению леммы 2 пересечениями этих α-уровневых множеств определяется множество α-соседних нечетких точек для заданной степени. Как видно из формулы (7), с уменьшением степени $\alpha = T(A, B)$ радиус этих окружностей увеличивается.

Понятно, что если алгоритмом FJP получено некоторое разбиение множества X, то выполняется неравенство

$$d_{\max}^{in} < d_{\min}^{out} < d_{\max}^{out}$$

и нет таких точек x, $y \in X$, чтобы выполнялось неравенство

$$d_{\max}^{in} < d(x, y) < d_{\min}^{out}.$$

Следовательно, в последовательности значений степеней α_i, $i = \overline{0, t}$, сгенерированных шагом 5 помехоустойчивого алгоритма FJP, некоторые последовательные значения α_z и α_{z+1} будут соответствовать значениям дистанций $d_{\max}^{in}$ и $d_{\min}^{out}$. Значения степеней отношения T, соответствующие значениям расстояний $0 = d_0$, $d_{\max}^{in}$, $d_{\min}^{out}$ и $d_{\max}^{out}$, обозначим через $1 = \alpha^0, \alpha^1, \alpha^2$ и α^3 соответственно.

Согласно шагам 8–10 для того чтобы алгоритм FJP выдавал предполагаемое

выше разбиение в качестве оптимального, должно выполняться равенство

$$\alpha^1 - \alpha^2 = \Delta\alpha_z = \alpha_z - \alpha_{z+1} = \max_{i=0,\,t-1} \Delta\alpha_i .$$

Предположим, что неравенства (16) выполняются. Сначала рассмотрим первую часть этих неравенств. Пусть удовлетворяется неравенство $\frac{d^{in}_{\max}}{d^{out}_{\min}} < \frac{1}{2}$. Отсюда, учитывая $d_0 = 0$, можно записать следующую цепочку неравенств:

$$2d^{in}_{\max} < d^{out}_{\min} \to d^{in}_{\max} < d^{out}_{\min} - d^{in}_{\max} \to d^{in}_{\max} - d_0 < d^{out}_{\min} - d^{in}_{\max}. \qquad (17)$$

С учетом отношения (14) имеем

$$\alpha^0 - \alpha^1 = \left(1 - \frac{d_0}{d_{\max}}\right) - \left(1 - \frac{d^{in}_{\max}}{d_{\max}}\right) = \frac{d^{in}_{\max} - d_0}{d_{\max}}$$

и

$$\alpha^1 - \alpha^2 = \left(1 - \frac{d^{in}_{\max}}{d_{\max}}\right) - \left(1 - \frac{d^{out}_{\min}}{d_{\max}}\right) = \frac{d^{out}_{\min} - d^{in}_{\max}}{d_{\max}}. \qquad (18)$$

Согласно (17) получаем, что

$$\alpha^0 - \alpha^1 < \alpha^1 - \alpha^2.$$

Тогда $\#\alpha_i, \alpha_j \in [\alpha^1, \alpha^0]$ справедливо неравенство

$$|\alpha_i - \alpha_j| \le \alpha^0 - \alpha^1 < \alpha^1 - \alpha^2.$$

Следовательно, $\#\alpha_i, \alpha_{i+1} \in [\alpha^1, \alpha^0]$ справедливо неравенство

$$\alpha_i - \alpha_{i+1} < \alpha^1 - \alpha^2. \qquad (19)$$

Далее рассмотрим вторую часть неравенства (16). Пусть выполняется неравенство

$$\frac{1}{2} < \frac{d^{out}_{\min} - d^{in}_{\max}}{d^{out}_{\max}}.$$

Отсюда вытекает

$$d^{out}_{\max} < 2(d^{out}_{\min} - d^{in}_{\max}).$$

Следовательно,

$$d^{out}_{\max} - d^{out}_{\min} < d^{out}_{\min} - 2d^{in}_{\max} < d^{out}_{\min} - d^{in}_{\max},$$

т.е.

$$d^{out}_{\max} - d^{out}_{\min} < d^{out}_{\min} - d^{in}_{\max}. \qquad (20)$$

Переходя на отношение T, можно записать

$$\alpha^2 - \alpha^3 = \left(1 - \frac{d^{out}_{\min}}{d_{\max}}\right) - \left(1 - \frac{d^{out}_{\max}}{d_{\max}}\right) = \frac{d^{out}_{\max} - d^{out}_{\min}}{d_{\max}}. \qquad (21)$$

Учитывая равенства (21) и (18), из неравенства (20) получаем $\alpha^2 - \alpha^3 < \alpha^1 - \alpha^2$. Тогда $\forall\alpha_i, \alpha_j \in [\alpha^3, \alpha^2]$ справедливо неравенство

$$|\alpha_i - \alpha_j| \le \alpha^2 - \alpha^3 < \alpha^1 - \alpha^2.$$

Следовательно, $\#\alpha_i, \alpha_{i+1} \in [\alpha^3, \alpha^2]$ справедливо неравенство

$$\alpha_i - \alpha_{i+1} < \alpha^1 - \alpha^2. \tag{22}$$

Таким образом, из (19) и (22) следует, что

$$\alpha^1 - \alpha^2 = \max_{i=0,\,t-1} \Delta\alpha_i.$$

Этим завершается доказательство теоремы.

Отметим, что доказанная выше теорема имеет непосредственное отношение к простому алгоритму FJP, предложенному в [6]. Для помехоустойчивого варианта алгоритма FJP, предложенного в настоящей статье, условия теоремы 2 должны выполняться только для точек ядер классов.

4. ВЫЧИСЛИТЕЛЬНЫЕ ЭКСПЕРИМЕНТЫ

Поведение помехоустойчивости алгоритма FJP и корректность его работы будем демонстрировать на базах данных с различными конфигурациями, представленными в [13]. Используемые алгоритмы запрограммированы на языке C++ для персональных компьютеров типа Pentium IV и проведены вычислительные эксперименты. На рис. 4–6 отражены результаты работы программ, где точки ядер классов показаны цифрами 0, 1, 2, а точки шума — более жирными квадратами. Так как классы смешаны множествами точек помеха, работа алгоритма FJP без механизма помехоустойчивости (см. [6]) будет давать некорректные результаты. Как видно из рис. 4–6, результат предложенного в настоящей статье помехоустойчивого варианта алгоритма FJP является корректным. Отметим, что при работе алгоритма сначала определяются ядра классов, а затем каждая точка шума передается ближайшему ядру.

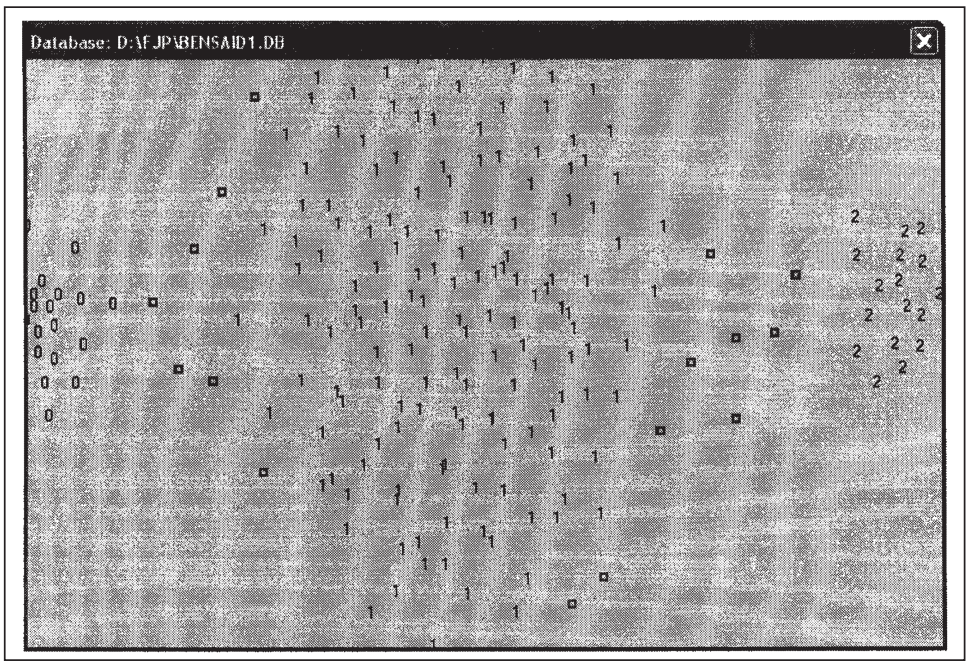

Рис. 4. Результат кластеризации БД-1 с параметрами $\varepsilon_1 = 0{,}9$, $\varepsilon_2 = 0{,}2$, где получено оптимальное значение $\overline{\alpha} = 0{,}9208$

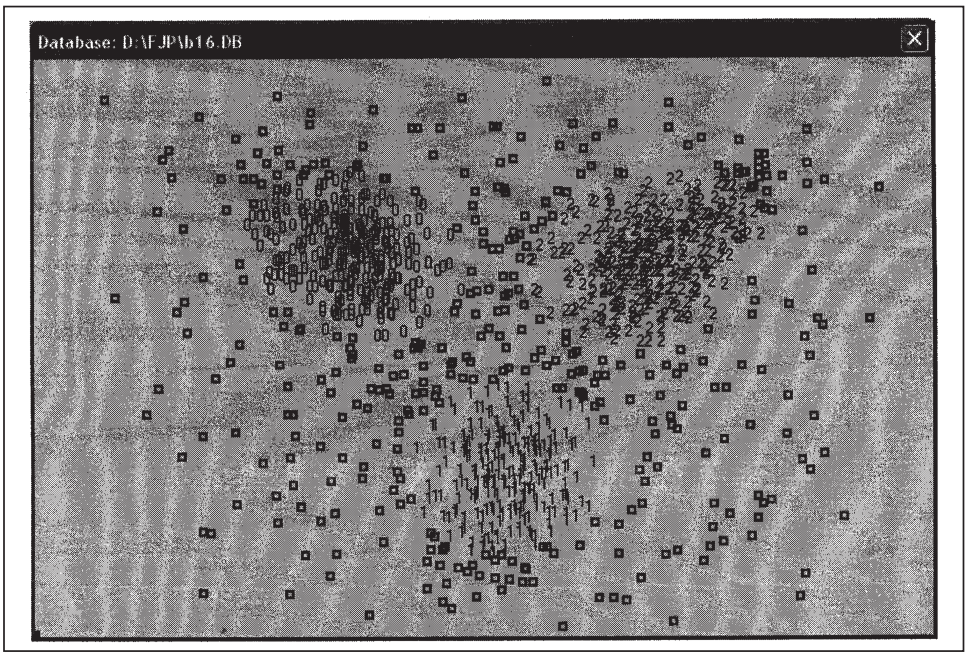

Рис. 5. Результат кластеризации БД-2 с параметрами $\varepsilon_1 = 0{,}9$, $\varepsilon_2 = 0{,}45$, где получено оптимальное значение $\overline{\alpha} = 0{,}9787$

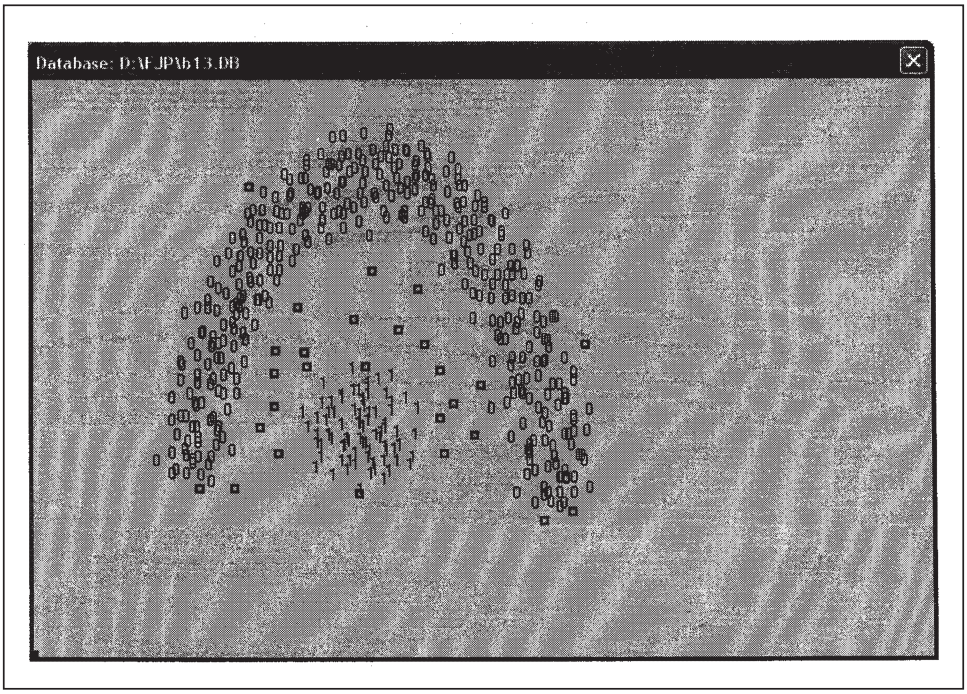

Рис. 6. Результат кластеризации БД-2 с параметрами $\varepsilon_1 = 0{,}9$, $\varepsilon_2 = 0{,}3$, где получено оптимальное значение $\overline{\alpha} = 0{,}9405$

Данные с несбалансированным распределением (БД-1). Вначале рассмотрим базу данных с классами, существенно различающимися по количеству точек. Результат кластеризации помехоустойчивым алгоритмом FJP с параметрами $\varepsilon_1 = 0.9$, $\varepsilon_2 = 0.2$, где получено оптимальное значение $\overline{\alpha} = 0.9208$, фиксирует рис. 4.

Данные с эллиптическими распределениями (БД-2). На рис. 5 отражена база данных, где классы имеют приблизительно одинаковые размеры с эллиптическими формами и приведен результат работы алгоритма с параметрами $\varepsilon_1 = 0.9, \varepsilon_2 = 0.45$, где получено оптимальное значение $\overline{\alpha} = 0.9787$.

Данные с линейно неразделимым распределением (БД-3). Рассмотрим базу данных с конфигурацией, где классы не могут быть разделены прямой линией. Результат работы помехоустойчивого алгоритма FJP для БД-3 с параметрами $\varepsilon_1 = 0.9, \varepsilon_2 = 0.3$, где получено оптимальное значение $\overline{\alpha} = 0.9465$, отражен на рис. 6.

Визуальное наблюдение расположения данных на рис. 4–6 подтверждает справедливость количества классов и их структуру, распознаваемых алгоритмом FJP. Отметим, что, как правило, классический алгоритм кластеризации FCM (Fuzzy c-Means) дает неадекватные результаты при кластеризации баз данных типа БД-1 и БД-3 [14].

ЗАКЛЮЧЕНИЕ

В настоящей статье был предложен помехоустойчивый вариант алгоритма FJP. Как было показано, основным достоинством предложенного алгоритма является возможность регулировки помехоустойчивости кластеризации и автоматического определения количества классов. Были использованы понятия нечеткой точки конусной формы, нечетких α-соседних и нечетких α-связанных точек, проведены исследования некоторых свойств относительно этих понятий. Доказано достаточное условие правильного распознавания алгоритмом FJP имеющейся скрытой структуры расположения классов при кластеризации.

Отметим, что рассмотренный алгоритм может быть использован как в качестве подготовительного этапа классического алгоритма нечеткой кластеризации FCM для определения начальных классов и их количества, так в качестве самостоятельного алгоритма кластеризации и распознавания образов.

СПИСОК ЛИТЕРАТУРЫ

1. Han J., Kamber M. Data mining: concepts and techniques. — San Diego: Acad. Press, 2001. — 550 p.
2. Dumitrescu D., Lazzerini B., Jain L.C. Fuzzy sets and their application to clustering and training. — New York: CRC Press LLC, 2000. — 622 p.
3. Насибов Э.Н. Идентификация состояний сложных систем с оценкой допустимой погрешности измерений при нечеткой информации // Кибернетика и системный анализ. — 2002. — **38**, № 1. — С. 63–71.
4. Насибов Э.Н. Об одной задаче идентификации состояний системы по нечетким значениям информативных признаков // Автоматика и вычисл. техника. — 2002. — **36**, № 2. — С. 33–40.
5. Nasibov E.N., Ulutagay G. Program tools for fuzzy clustering analysis // Proc. Intern. Scientific Conf. «Inform. Tech. and Telecomm. in Educ. and Science». — Antalya (Turkey), 2006. — P. 195–197.
6. Насибов Э.Н., Улутагай Г. Новый подход к проблеме кластеризации с использованием метода нечетких связанных точек // Автоматика и вычисл. техника. — 2005. — **39**, № 6. — С. 11–21.
7. Pal N.R., Bezdek J.C. On cluster validity for the fuzzy C-means model // IEEE Trans. on Fuzzy Systems. — 1995. — **3**, N 3. — P. 370–379.
8. Zahid N., Abouelala O., Limouri M., Essaid A. Fuzzy clustering based on *K*-near-

est-neighbours rule // Fuzzy Sets and Systems. — 2001. — **120**. — P. 239–247.

9. Velthuizen R.P., Hall L.O., Clarke L.P., Silbiger M.L. An investigation of mountain method clustering for large data sets // Pattern Recognition. — 1997. — **30,** N 7. — P. 1121–1135.
10. Yager R.R., Filev D.P. Approximate clustering via the mountain method // IEEE Trans. on Systems, Man and Cybernetics. — 1994. — **24**, N 8. — P. 1279–1284.
11. Dunn J.C. A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters // Cybernetics. — 1973. — **3**, N 3. — P. 32–57.
12. Hammah R.E., Curran J.H. On distance measures for the fuzzy *K*-means algorithm for joint data // Rock Mech. Rock Eng. — 1999. — **32**, N 1. — P. 1–27.
13. Tao C.W. Unsupervised fuzzy clustering with multi-center clusters // Fuzzy Sets and Systems. — 2002. — **128**. — P. 305–322.
14. Bensaid A.M., Hall L.O., Bezdek J.C. et al. Validity-Guided (Re)clustering with applications to image segmentation // IEEE Trans. on Fuzzy Systems. — 1996. — **4**, N 2. — P. 112–123.

Поступила 08.02.2006