ИНФОРМАТИКА

2009 **октябрь-декабрь** **№ 4**

УДК 519.237.8; 510.22

К.М. Садовская

АНАЛИЗ УСТОЙЧИВОСТИ МЕТОДОВ НЕЧЕТКОЙ КЛАСТЕРИЗАЦИИ К АНОМАЛЬНЫМ НАБЛЮДЕНИЯМ

Исследуется проблема устойчивости решений задачи нечеткой кластеризации по отношению к включению в исследуемую совокупность аномальных наблюдений. Рассматриваются целевые функционалы распространенных оптимизационных методов нечеткой кластеризации FCM, NC, PCM, FRC.

Введение

В последние годы в области нечеткой кластеризации обнаруживается тенденция к разработке методов, обеспечивающих устойчивость решений в условиях возможного наличия в исследуемой совокупности аномальных наблюдений. Под аномальными следует понимать наблюдения, принадлежащие классам, число представителей которых в исследуемой совокупности существенно мало в сравнении с числом представителей основных классов.

Наличие аномальных наблюдений часто не позволяет исследователю точно определить количество классов, представляющих исследуемую совокупность, а также сделать предварительную оценку плотности их распределения, что, безусловно, сказывается на результатах кластеризации.

При решении задачи нечеткой кластеризации с целью устранения существенного влияния аномальных наблюдений применяется либо их исключение из выборки на этапе предобработки, либо оптимизационный поиск решений с помощью целевого функционала, ослабляющего влияние аномальных наблюдений на определение принадлежности наблюдений основным классам.

Существование многообразия функционалов, построенных для ослабления влияния аномальных наблюдений на результаты кластеризации, обусловливает необходимость их анализа с целью осмысленного выбора наиболее подходящего метода кластеризации.

Задача нечеткой кластеризации

Под решением задачи нечеткой кластеризации будем понимать разбиение P исследуемой совокупности наблюдений X на заданное число c нечетких множеств $A^i, i = 1,\ldots,c$. Под нечетким множеством (кластером) A^i понимается совокупность наблюдений $X = \{x_1,...,x_n\}$ с заданной на ней функцией принадлежности наблюдений данному множеству $\mu_{ij} = \mu_i(x_j)$, $j = 1,\ldots,n$, значения которой удовлетворяют условию $\mu_{ij} \in [0,1]$. Считаем, что нечеткие кластеры A^i, определенные на множестве наблюдений X, с соответствующими функциями принадлежности μ_{ij}, $i = 1,\ldots,c$, $j = 1,\ldots,n$, образуют нечеткое c-разбиение $P = \{A^1,\ldots, A^c\}$, если для каждого наблюдения $x_j \in X$ выполняется условие

$$\sum_{i=1}^{c} \mu_{ij} = 1. \qquad (1)$$

Задача нечеткой кластеризации заключается в нахождении условного экстремума целевого функционала $F = F(P)$ на множестве Π всех нечетких c-разбиений P.

Рассмотрим обобщенный вид целевого функционала, представляющий собой суммарное взвешенное отклонение наблюдений $X = \{x_1,..., x_n\}$ от центров тяжести кластеров $\{\tau_1,..., \tau_c\}$ (центроидов):

$$F(P)=\sum_{i=1}^{c}\sum_{j=1}^{n} f\left(\mu_{ij}\right) d_{ij}\,, \tag{2}$$

где $d_{ij}=d(x_j,\tau_i)$ – функция расстояния от наблюдения x_j до центроида τ_i нечеткого кластера $A^i\in\{A^1,\ldots,A^c\}$; $\mu_{ij}=\mu_i(x_j)$ – функция принадлежности наблюдения $x_j\in X$ нечеткому кластеру A^i.

Основой для разработки большинства оптимизационных методов нечеткой кластеризации послужил целевой функционал Дж. Данна и Дж. Беждека [1]:

$$F_{FCM}(P)=\sum_{i=1}^{c}\sum_{j=1}^{n}\mu_{ij}^{\gamma}d_{ij}^{2}\,, \tag{3}$$

где $d_{ij}=d(x_j,\tau_i)$ – функция расстояния, в качестве которой исследователями, как правило, используется функция евклидова расстояния в m-мерном признаковом пространстве R^m: $d_{ij}=\left\|x_j-\tau_i\right\|=\sqrt{\sum_{k=1}^{m}(x_{jk}-\tau_{ik})^2}$; γ – показатель нечеткости кластеризации, $1<\gamma<\infty$.

В рамках задачи кластеризации требуется определить такое оптимальное нечеткое c-разбиение $P=\{A^1,\ldots,A^c\}$, для которого оценка (3) принимает наименьшее значение.

Необходимые условия экстремума функционала (3) могут быть представлены системой уравнений

$$\begin{cases}\mu_{ij}=\dfrac{1}{\sum\limits_{k=1}^{c}\left[\dfrac{d_{ji}^2}{d_{jk}^2}\right]^{1/(\gamma-1)}},\ i=1,\ldots,c,\ j=1,\ldots,n;\\ \tau_v=\dfrac{\sum\limits_{t=1}^{n}\mu_{vt}^{\gamma}x_t}{\sum\limits_{t=1}^{n}\mu_{vt}^{\gamma}},\quad v=1,\ldots,c,\ t=1,\ldots,n.\end{cases} \tag{4}$$

Решение оптимизационной задачи может быть найдено в соответствии с алгоритмом FCM (нечетких *c*-средних) [1].

Задача нечеткой кластеризации в условиях аномальных наблюдений

Предположим, что исследуемая совокупность содержит наблюдения непредставительных классов, число которых не может быть определено априори. Предположение о наличии в исследуемой совокупности аномальных наблюдений, т. е. наблюдений, которые могут принадлежать неосновным классам, приводит к ослаблению условия (1):

$$\sum_{i=1}^{c}\mu_{ij}\le 1\,, \tag{5}$$

где c – число основных классов, устанавливаемое до процедуры кластеризации.

Замена условия (1) условием (5) влечет за собой необходимость переопределения задачи минимизации целевого функционала. Так, для целевого функционала F_{FCM}, заданного выражением (3), при условии (5) существует вырожденное решение задачи кластеризации: $\mu_{ij}=0$, $i=1,\ldots,c,\ j=1,\ldots,n$.

Обзор ряда методов кластеризации оптимизационного направления, нацеленных на обеспечение устойчивости решения задачи нечеткой кластеризации при наличии в исследуемой совокупности аномальных наблюдений, приведен в работе [2]. Указанные в работе методы нечеткой кластеризации NC [3], PCM [4] и FRC [5] представляют собой эволюцию способов решения задачи нечеткой кластеризации в условиях аномальных наблюдений. Метод NC предусматривает ослабление влияния аномального наблюдения на совокупность основных кластеров в целом, PCM и FRC – на каждый кластер в отдельности, нивелируя в том числе и влияние наблюдений, находящихся в межкластерном пространстве.

Методы нечеткой кластеризации NC, PCM и FRC основаны на построении целевого функционала как суммы оценки разбиения исследуемой совокупности наблюдений на множестве основных кластеров $\{A^1,\ldots, A^c\}$, которая представима в виде (3), и оценки разбиения исследуемой совокупности на дополнительные нечеткие множества $\{B^1,\ldots, B^{c^*}\}$, содержащие аномальные наблюдения классифицируемой выборки:

$$F_{NC}(P) = \sum_{i=1}^{c}\sum_{j=1}^{n}\mu_{ij}^{\gamma} d_{ij}^2 + \sum_{j=1}^{n}\delta^2(1-\sum_{i=1}^{c}\mu_{ij})^{\gamma}\ ; \tag{6}$$

$$F_{PCM}(P) = \sum_{i=1}^{c}\sum_{j=1}^{n}\mu_{ij}^{\gamma} d_{ij}^2 + \sum_{i=1}^{\tilde{n}}\delta_i^2\sum_{j=1}^{n}(1-\mu_{ij})^{\gamma}\ ; \tag{7}$$

$$F_{FRC}(P) = \sum_{i=1}^{c}\sum_{j=1}^{n}\mu_{ij} d_{ij}^{2p} + \sum_{i=1}^{c}\sum_{j=1}^{n}\left(1+\mu_{ij} * \log(\mu_{ij}) - \mu_{ij}\right)\delta_i^{2p}\ , \tag{8}$$

где $P = \{A^1,\ldots, A^c, B^1,\ldots, B^{c^*}\}$; $d_{ij} = d(x_j, \tau_i)$ – функция расстояния; δ, δ_i – параметры, характеризующие дисперсию всей исследуемой совокупности и внутриклассовую дисперсию кластера A^i соответственно; γ, p – показатели нечеткости кластеризации. Выбор функции $d_{ij} = d(x_j, \tau_i)$ и значений параметров δ, δ_i, γ, p для классификации произвольной совокупности является открытым вопросом нечеткой кластеризации. Примеры предложений по выбору упомянутых параметров кластеризации содержатся в работах [3–7].

Для исследования проблемы устойчивости решений задачи нечеткой кластеризации в условиях возможного наличия аномальных наблюдений в исследуемой совокупности построим функцию относительного влияния произвольного наблюдения, в частности аномального, на оценку заданного разбиения и изучим ее свойства.

Заметим, что целевые функционалы, заданные выражениями (3), (6)–(8), представимы в общем виде как сумма независимых компонент, каждая из которых зависит только от взаимного расположения произвольного наблюдения x_j и совокупности центроидов основных кластеров $\{\tau^i\}_{i=1}^{c}$:

$$F(P) = \sum_{j=1}^{n} Q(x_j)\ ,$$

где $Q_{FCM}(x) = \sum_{i=1}^{c}\mu_i^{\gamma}(x) d^2(x,\tau_i)$;

$$Q_{NC}(x) = \sum_{i=1}^{c}\mu_i^{\gamma}(x) d^2(x,\tau_i) + \delta^2(1-\sum_{i=1}^{c}\mu_i(x))^{\gamma}\ ;$$

$$Q_{PCM}(x) = \sum_{i=1}^{c}\mu_i^{\gamma}(x) d^2(x,\tau_i) + \delta_i^2(1-\mu_i(x))^{\gamma}\ ;$$

$$Q_{FRC}(x) = \sum_{i=1}^{c}\mu_i(x) d^{2p}(x,\tau_i) + \left(1+\mu_i(x) * \log(\mu_i(x)) - \mu_i(x)\right)\delta_i^{2p}\ .$$

Таким образом, целевой функционал может быть охарактеризован показательной функцией $Q = Q(x)$, значение которой в произвольной точке x является некоторой долей, вносимой наблюдением x в общую оценку разбиения P при условии принадлежности x исследуемой совокупности.

Рассмотрим свойства показательных функций $Q_{FCM}(x)$, $Q_{NC}(x)$, $Q_{PCM}(x)$, $Q_{FRC}(x)$ на примере задач кластеризации в общей постановке с одним и двумя основными кластерами. Сравнительный анализ показательных функций будем проводить, используя в качестве функции расстояния функцию евклидова расстояния в m-мерном признаковом пространстве R^m:

$$d_{ij} = \left\| x_j - \tau_i \right\| = \sqrt{\sum_{k=1}^{m} (x_{jk} - \tau_{ik})^2} \ .$$

Задача кластеризации с одним основным кластером

Рассмотрим задачу кластеризации наблюдений в m-мерном признаковом пространстве R^m с одним основным кластером A. Взаимное расположение произвольного наблюдения $x \in \mathrm{R}^m$ и центроида τ кластера A может быть охарактеризовано скалярным параметром расстояния d между точками x и τ. Поэтому показательная функция $Q(x)$ может быть рассмотрена как функция вида $Q = Q(d)$.

Проведем анализ показательных функций:

$$Q^I_{FCM}(d) = \mu^\gamma d^2; \tag{9}$$

$$Q^I_{NC}(d) = \mu^\gamma d^2 + (1-\mu\ )^\gamma \delta^2; \tag{10}$$

$$Q^I_{PCM}(d) = \mu^\gamma d^2 + (1-\mu)^\gamma \delta^2; \tag{11}$$

$$Q^I_{FRC}(d) = \mu d^{2p} + (1 + \mu * \log(\mu) - \mu)\delta^{2p}, \tag{12}$$

где $d = d(x,\tau) = \left\| \tau - x \right\|$ – расстояние от наблюдения x до центроида τ; $\mu = \mu^A(x)$ – степень принадлежности наблюдения x кластеру A, которая в соответствии с необходимыми условиями экстремума целевого функционала может быть определена как функция от расстояния d:

$$\mu_{FCM}(d) = 1; \tag{13}$$

$$\mu_{NC}(d) = \left(1 + \left(\frac{d^2}{\delta^2} \right)^{\frac{1}{\gamma - 1}} \right)^{-1}; \tag{14}$$

$$\mu_{PCM}(d) = \left(1 + \left(\frac{d^2}{\delta^2} \right)^{\frac{1}{\gamma - 1}} \right)^{-1}; \tag{15}$$

$$\mu_{FRC}(d) = \exp\left(-\left(\frac{d^2}{\delta^2} \right)^{p} \right). \tag{16}$$

Заметим, что соотношения (10), (11) определяют одно функциональное выражение. Поэтому в дальнейшем при исследовании показательных функций в условиях задачи с одним основным кластером не будем различать функции $Q^I_{NC}(d)$ и $Q^I_{PCM}(d)$, обозначив их как $Q^I_{NC,PCM}(d)$. Также в дальнейшем под обозначением $Q^I(d)$ будем понимать совокупность показательных функций $Q^I_{FCM}(d)$, $Q^I_{NC}(d)$, $Q^I_{PCM}(d)$, $Q^I_{FRC}(d)$, заданных выражениями (9)–(12), в случае, если они обладают каким-либо общим свойством.

На рис. 1 изображены показательные функции $Q^I(d)$ для характерных расчетных значений параметров:

$$\delta = 0{,}5;\ \mathrm{p} = 1;\ \gamma = 2. \tag{17}$$

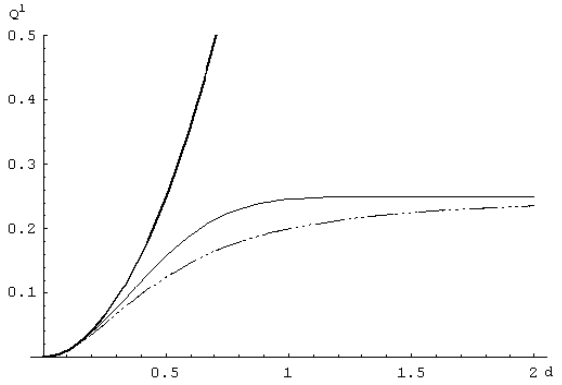

Рис. 1. График показательных функций $Q^I_{FCM}(d)$ – «—», $Q^I_{NC,PCM}(d)$ – «– · · –», $Q^I_{FRC}(d)$ – «– » для расчетных значений параметров (17)

Проведем элементарный сравнительный функциональный анализ показательных функций $Q^I(d)$. Функции $Q^I(d)$, определенные на множестве неотрицательных действительных чисел R_+, непрерывны и дифференцируемы во всей области определения, имеют единственный нуль в точке $d = 0$ ($x_0 = \tau$). Асимптоты показательных функций:

$$\lim_{d\to\infty} Q^I_{FCM}(d) = \infty,\ \lim_{d\to\infty} Q^I_{NC,PCM}(d) = \delta^2,\ \lim_{d\to\infty} Q^I_{FRC}(d) = \delta^2 \text{ при } \delta \in (0;\infty). \tag{18}$$

Дифференцируя выражения (9)–(12) по переменной d с учетом функциональной зависимости параметра степени принадлежности μ от d (13)–(16), получим выражения для первых производных $Q^{I(1)}(d)$ показательных функций $Q^I(d)$:

$$Q^{I(1)}_{FCM}(d) = 2d; \tag{19}$$

$$Q^{I(1)}_{NC,PCM}(d) = 2d\left(1+\left(\frac{d^2}{\delta^2}\right)^{\frac{1}{\gamma-1}}\right)^{-\gamma}; \tag{20}$$

$$Q^{I(1)}_{FRC}(d) = 2pd^{2p-1}\exp\left(-\left(\frac{d^2}{\delta^2}\right)^{p}\right). \tag{21}$$

Заметим, что функции $Q^I(d)$ имеют единственный экстремум в точке $d=0$ и являются монотонно возрастающими на всем множестве $(0,+\infty)$.

Общие выражения для вторых производных показательных функций $Q^I(d)$:

$$Q^{I(2)}_{FCM}(d)=2; \tag{22}$$

$$Q^{I(2)}_{NC,PCM}(d)=2\left(1+\left(\frac{d^2}{\delta^2}\right)^{\frac{1}{\gamma-1}}\right)^{-\gamma-1}\left(1+\frac{(1+\gamma)}{(1-\gamma)}\left(\frac{d^2}{\delta^2}\right)^{\frac{1}{\gamma-1}}\right); \tag{23}$$

$$Q^{I(2)}_{FRC}(d)=-2p\left(1+2p\left(-1+\left(\frac{d^2}{\delta^2}\right)^{p}\right)\right)d^{2p-2}\exp\left(-\left(\frac{d^2}{\delta^2}\right)^{p}\right). \tag{24}$$

Функция $Q^{I(2)}_{FCM}(d)$ не имеет нулей. Нули вторых производных функций $Q^{I}_{NC,PCM}(d)$ и $Q^{I}_{FRC}(d)$ могут быть выражены аналитически как функции параметров δ, γ, p:

$$Q^{I(2)}_{NC,PCM}(d_1)=0 \Leftrightarrow d_1=\delta\sqrt{\left(\frac{\gamma-1}{1+\gamma}\right)^{\gamma-1}}; \tag{25}$$

$$Q^{I(2)}_{FRC}(d_1)=0 \Leftrightarrow d_1=\delta\left(1-\frac{1}{2p}\right)^{\frac{1}{2p}}. \tag{26}$$

Значения d_1 для характерных расчетных значений параметров (17) (рис. 2): $Q^{I(2)}_{NC,PCM}(d_1)=0 \Leftrightarrow d_1=\delta/\sqrt{3}$, $Q^{I(2)}_{FRC}(d_1)=0 \Leftrightarrow d_1=\delta/\sqrt{2}$.

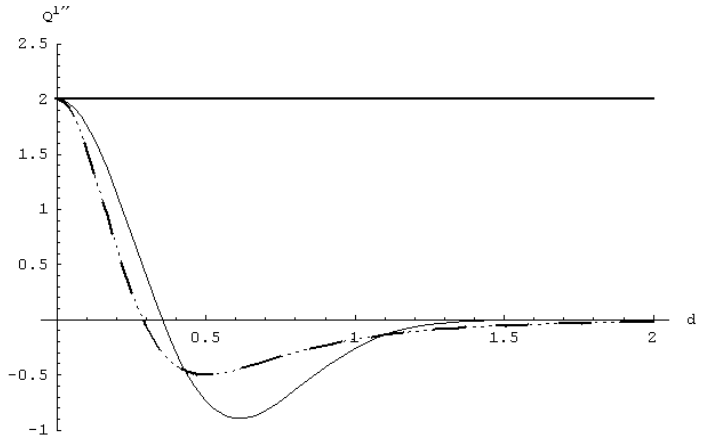

Рис. 2. График второй производной функций $Q^{I}_{FCM}(d)$ – «—», $Q^{I}_{NC,PCM}(d)$ – «– · – », $Q^{I}_{FRC}(d)$ – «– » для расчетных значений параметров (17)

Пользуясь результатами сравнительного анализа показательных функций $Q^I(d)$, заданных выражениями (9)–(12), можно описать характерные особенности целевых функционалов алгоритмов FCM, NC, PCM, FRC.

Для показательных функций $Q^I(d)$ и соответственно целевых функционалов оптимальным условием взаиморасположения центроида основного кластера A и произвольного наблюдения x исследуемой совокупности является их точное совпадение. Это очевидное утверждение подтверждается тем фактом, что глобальный минимум функций $Q^I(d)$ достигается в точке $d = 0$.

Характер монотонного возрастания функций $Q^I(d)$ во всей области их определения, за исключением точки $d = 0$, свидетельствует о прямой зависимости доли, вносимой наблюдением x в общую оценку нечеткого разбиения P, от расстояния между x и центроидом основного кластера A.

Основным отличительным свойством показательных функций $Q^I_{NC,PCM}(d)$ и $Q^I_{FRC}(d)$ необходимо считать их ограниченность при больших значениях аргумента, которая свидетельствует об ограниченности влияния аномальных наблюдений на оценку разбиения и соответственно на результаты кластеризации.

Для области малых значений аргумента функции $Q^I_{NC,PCM}(d)$ и $Q^I_{FRC}(d)$ характеризуются увеличением скорости роста с ростом значения аргумента. Для больших значений аргумента функции замедляют свой рост и согласно (18) имеют предельное значение δ^2. Область малых значений аргумента показательных функций $Q^I_{NC,PCM}(d)$ и $Q^I_{FRC}(d)$ может трактоваться как некоторая доверительная область основного кластера, а область больших значений аргумента, где доля, вносимая наблюдением в общую оценку разбиения, не зависит существенно от расстояния между наблюдением и центроидом, – как область аномальных наблюдений. Граница указанных областей согласно выражениям (25), (26) непосредственно зависит от значений параметров δ, γ и p.

Все рассмотренные выше свойства показательных функций характерны для задачи, описывающей вырожденную проблему кластеризации наблюдений с одним основным кластером. Решение задач кластеризации в общем случае может быть рассмотрено как суперпозиция решений вырожденной проблемы кластеризации исследуемой совокупности относительно каждого основного кластера в отдельности.

Задача кластеризации с двумя основными кластерами

Рассмотрим задачу кластеризации наблюдений в m-мерном признаковом пространстве R^m для случая двух основных кластеров A^1 и A^2. Пусть расстояние между центроидами τ_1 и τ_2 соответствующих кластеров A^1 и A^2 задано как параметр $s = d(\tau_1, \tau_2)$. Систему координат выберем таким образом, чтобы центроид τ_1 совпал с началом отсчета и τ_2 принадлежал оси абсцисс.

Исследуем показательные функции

$$Q^{II}_{FCM}(x) = \sum_{i=1}^{c} \mu_i^{\gamma}(x) d^2(\tau_i, x); \tag{27}$$

$$Q^{II}_{NC}(x) = \sum_{i=1}^{c} \mu_i^{\gamma}(x) d^2(\tau_i, x) + (1 - \sum_{i=1}^{c} \mu_i(x))^{\gamma} \delta^2; \tag{28}$$

$$Q^{II}_{PCM}(x) = \sum_{i=1}^{c} \mu_i^{\gamma}(x) d^2(\tau_i, x) + \sum_{i=1}^{c} (1 - \mu_i(x))^{\gamma} \delta_i^2; \tag{29}$$

$$Q^{II}_{FRC}(x) = \sum_{i=1}^{c} \mu_i(x) d^{2p}(\tau_i, x) + \sum_{i=1}^{c} (1 + \mu_i(x) * \log(\mu_i(x)) - \mu_i(x)) \delta_i^{2p}, \tag{30}$$

где $x \in \mathrm{R}^m$ – вектор наблюдения; $c = 2$ – число основных кластеров; $d(\tau_i, x) = \|\tau_i - x\|$ – расстояние от наблюдения x до центроида τ_i, $i = 1,2$; $\mu_i(x) = \mu_{A^i}(x)$ – степень принадлежности

наблюдения x кластеру A^i, которая в соответствии с необходимыми условиями экстремума целевых функционалов может быть определена следующим образом:

$$\begin{cases} \mu_{FCM\,i}(x) = \dfrac{1}{\sum\limits_{k=1}^{2}\left[\dfrac{d^2(\tau_i,x)}{d^2(\tau_k,x)}\right]^{1/(\gamma-1)}},\ x \neq \tau_j,\ j=\overline{1,2}, \\ \mu_{FCM\,i}(x) = 1,\ x = \tau_i, \\ \mu_{FCM\,i}(x) = 0,\ x = \tau_j,\ j \neq i; \end{cases} \tag{31}$$

$$\begin{cases} \mu_{NCi}(x) = \dfrac{1}{\left(\dfrac{d^2(\tau_i,x)}{\delta^2}\right)^{\frac{1}{\gamma-1}} + \sum\limits_{k}^{2}\left(\dfrac{d^2(\tau_i,x)}{d^2(\tau_k,x)}\right)^{\frac{1}{\gamma-1}}},\ x \neq \tau_j,\ j=\overline{1,2}, \\ \mu_{NCi}(x) = 1,\ x = \tau_i, \\ \mu_{NCi}(x) = 0,\ x = \tau_j,\ j \neq i; \end{cases} \tag{32}$$

$$\mu_{PCM\,i}(x) = \frac{1}{1+\left(\dfrac{d^2(\tau_i,x)}{\delta_i^{\,2}}\right)^{\frac{1}{\gamma-1}}}; \tag{33}$$

$$\mu_{FRC\,i}(x) = \exp\left(-\left(\frac{d^2(\tau_i,x)}{\delta_i^{\,2}}\right)^{p}\right). \tag{34}$$

Под обозначением $Q^{II}(x)$ в дальнейшем будем понимать совокупность показательных функций $Q^{II}_{FCM}(x)$, $Q^{II}_{NC}(x)$, $Q^{II}_{PCM}(x)$, $Q^{II}_{FRC}(x)$, заданных выражениями (27)–(30) соответственно. На рис. 3 изображены показательные функции $Q^{II}(x)$ для одномерного случая $x \in R$ при следующих характерных расчетных значениях параметров:

$$s = 1;\ \delta = 0{,}5;\ \delta_1 = 0{,}5;\ \delta_2 = 0{,}5;\ \gamma = 2;\ p = 1. \tag{35}$$

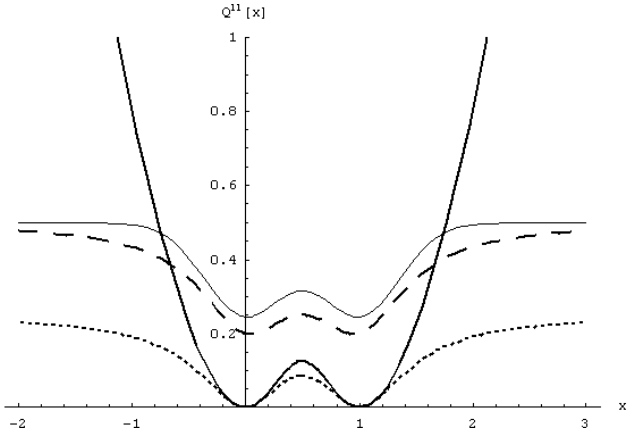

Рис. 3. График показательных функций $Q^{II}_{FCM}(x)$ – «—», $Q^{II}_{NC}(x)$ – «· · · ·», $Q^{II}_{PCM}(x)$ – «– – –», $Q^{II}_{FRC}(x)$ – «– · –» для одномерного признакового пространства при расчетных значениях параметров (35)

Проведем элементарный функциональный анализ показательных функций $Q^{II}(x)$. Функции $Q^{II}(x)$, определенные на множестве R^m, непрерывны и дифференцируемы во всей области определения. Функции $Q^{II}_{FCM}(x)$, $Q^{II}_{NC}(x)$ имеют нули в точках $x=\tau_1$, $x=\tau_2$; $Q^{II}_{PCM}(x)$, $Q^{II}_{FRC}(x)$ нулей не имеют. Асимптоты функций $Q^{II}(x)$ при $\delta\in(0;\infty)$:

$$\lim_{x\to\infty} Q^{II}_{FCM}(x)=\infty,\ \lim_{x\to\infty} Q^{II}_{NC}(x)=\delta^2,\ \lim_{x\to\infty} Q^{II}_{PCM}(x)=\sum_i^c \delta_i^2,\ \lim_{x\to\infty} Q^{II}_{FRC}(x)=\sum_i^c \delta_i^2 . \qquad (36)$$

Используя общие выражения (27)–(30) для показательных функций $Q^{II}_{FCM}(x)$, $Q^{II}_{NC}(x)$ и выражения (30)–(33) степеней принадлежностей как функций наблюдения x, получим

$$Q^{II}_{FCM}(\tau_i)=\sum_{k=1}^c \mu_k(\tau_i)^\gamma d(\tau_i,\tau_k)^2 = 1^\gamma 0^2 + 0^\gamma s^2 = 0;$$

$$Q^{II}_{NC}(\tau_i)=\sum_{k=1}^c \mu_k(\tau_i)^\gamma d(\tau_k,\tau_i)^2 + (1-\sum_{k=1}^c \mu_k(\tau_i))^\gamma \delta^2 = 1^\gamma 0^2 + 0^\gamma s^2 + (1-1-0)\delta^2 = 0.$$

Значит, точки $x=\tau_1$ и $x=\tau_2$ являются экстремумами неотрицательных показательных функций $Q^{II}_{FCM}(x)$ и $Q^{II}_{NC}(x)$.

Запишем выражения для первых производных показательных функций $Q^{II}_{FCM}(x)$ и $Q^{II}_{NC}(x)$, полученные в результате дифференцирования выражений (27), (28) с учетом функциональных зависимостей (31), (32):

$$Q^{II(1)}_{FCM}(x)=2\left(\sum_{k=1}^2 d_k d'_k \mu_k^\gamma + \frac{\gamma}{\gamma-1}\sum_{k=1}^2\left(\frac{d_k}{d_i}\right)^{\frac{\gamma+1}{\gamma-1}}\left(d'_k d_i - d_k d'_i\right)\mu_k^{\gamma+1}\right), \qquad (37)$$

где $\mu_k=\mu_k^{FCM}=\left(1+\left[\frac{d_k^2}{d_i^2}\right]^{1/(\gamma-1)}\right)^{-1}$;

$$Q^{II(1)}_{NC}(x)=2\sum_{k=1}^2 d_k d'_k \mu_k^\gamma + \frac{2\gamma}{\gamma-1}\sum_{k=1}^2 d_k d_k^{\frac{2}{\gamma-1}}\left(d'_k\left(\frac{1}{\delta^{\frac{2}{\gamma-1}}}+\frac{1}{d_i^{\frac{2}{\gamma-1}}}\right)-\frac{d_k d'_i}{d_i^{\frac{\gamma+1}{\gamma-1}}}\right)\left(\frac{\delta^2}{d_k^2}\left(1-\sum_{t=1}^2\mu_t\right)-\mu_k^{\gamma+1}\right), \qquad (38)$$

где $\mu_k=\mu_k^{NC}=\left(\left(\frac{d_k^2}{\delta^2}\right)^{\frac{1}{\gamma-1}}+\left[\frac{d_k}{d_i}^2\right]^{1/(\gamma-1)}+1\right)^{-1}$, $d'_k(x)=\frac{\sum_{j=1}^m (x^j-\tau_k^j)}{\sqrt{\sum_{j=1}^m (\tau_k^j-x^j)^2}}$.

Аналогично могут быть получены выражения для первых производных функций $Q^{II}_{PCM}(x)$ и $Q^{II}_{FRC}(x)$ после дифференцирования выражений (29), (30) с учетом (33), (34):

$$Q^{II(1)}_{PCM}(x)=\sum_{k=1}^2 2 d_k d'_k\left(1+\left(\frac{d_k^2}{\delta_k^2}\right)^{\frac{1}{\gamma-1}}\right)^{-\gamma};$$

$$Q_{FRC}^{II(1)}(x) = 2p\sum_{k=1}^{2}\exp\left(-\left(\frac{d_k^2}{\delta_k^2}\right)^p\right)\left(\frac{d_k^2}{\delta_k^2}\right)^p \delta_k^{2p} d_k^{-1} d_k',$$

где $d_k = d_k(x) = \sqrt{\sum_{j=1}^{m}(\tau_k^j - x^j)^2}$, $d_k'(x) = \dfrac{\sum_{j=1}^{m}(x^j - \tau_k^j)}{\sqrt{\sum_{j=1}^{m}(\tau_k^j - x^j)^2}}$.

Заметим, что функции $Q_{PCM}^{II}(x)$, $Q_{FRC}^{II}(x)$ не имеют экстремумов в точках τ_1 и τ_2:

$$Q_{PCM}^{II(1)}(\tau_i) = 2s\left(1+\left(\frac{s^2}{\delta_k^2}\right)^{\frac{1}{\gamma-1}}\right)^{-\gamma}\left((-1)^i + \frac{\gamma}{\gamma-1}\left(1+\left(\frac{s^2}{\delta_k^2}\right)^{\frac{1}{\gamma-1}}\right)^{-1}\left((-1)^{i+1}\left(\frac{s^2}{\delta_k^2}\right)^{\frac{1}{\gamma-1}} + (-1)^i\left(\frac{s^2}{\delta_k^2}\right)^{\frac{\gamma}{\gamma-1}}\right)\right) \neq 0;$$

$$Q_{FRC}^{II(1)}(\tau_i) = 2p\exp\left(-\left(\frac{s^2}{\delta_k^2}\right)^p\right)s^{2p}(-1)^i \neq 0.$$

На рис. 4 изображены функции первых производных $Q^{II(1)}(x)$ для одномерного случая при значениях параметров (35).

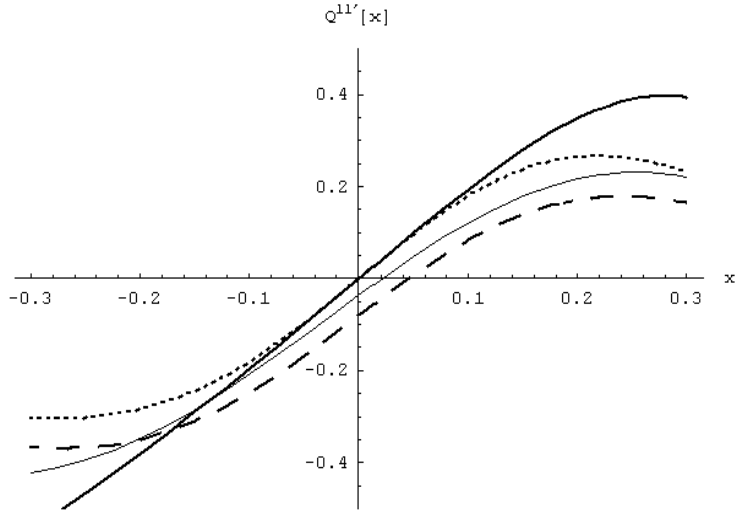

Рис. 4. График первой производной функций $Q_{FCM}^{II}(x)$ – «—», $Q_{NC}^{II}(x)$ – «· · · ·», $Q_{PCM}^{II}(x)$ – «– – –», $Q_{FRC}^{II}(x)$ – «– –» в окрестности центроида τ_1 для расчетных значений параметров (35)

Пользуясь результатами сравнительного анализа показательных функций $Q^{II}(x)$, опишем основные свойства целевых функционалов алгоритмов FCM, NC, PCM, FRC.

Для каждого основного кластера может быть выделена некоторая доверительная область, где показательные функции $Q_{FCM}^{II}(x)$, $Q_{NC}^{II}(x)$, $Q_{PCM}^{II}(x)$ или $Q_{FRC}^{II}(x)$ имеют строгий локальный минимум. Необходимо заметить, что экстремальные значения функций $Q_{FCM}^{II}(x)$ и $Q_{NC}^{II}(x)$ достигаются в точках центроидов, но для функций $Q_{PCM}^{II}(x)$ и $Q_{FRC}^{II}(x)$ это утверждение неверно. Невыполнение условий экстремума в точках центроидов обусловлено взаимным влиянием оценок распределения основных кластеров. Анализ показательных функций объясня-

ет этот неочевидный факт, известный многим исследователям из опыта практического применения методов нечеткой кластеризации.

Границей доверительной области основного кластера будем считать точки нулей второй производной показательной функции. Доверительная область основных кластеров функций $Q^{II}_{NC}(x)$, $Q^{II}_{PCM}(x)$ и $Q^{II}_{FRC}(x)$ всегда ограничена.

За пределами доверительных областей основных кластеров при удалении наблюдения от любого из центроидов отмечено монотонное возрастание функций $Q^{II}_{NC}(x)$, $Q^{II}_{PCM}(x)$ и $Q^{II}_{FRC}(x)$. Асимптотическое поведение показательных функций $Q^{II}_{NC}(x)$, $Q^{II}_{PCM}(x)$ и $Q^{II}_{FRC}(x)$ обусловливает незначительное изменение доли, вносимой аномальными наблюдениями в оценку разбиения в случае изменения положения центроидов в области основных кластеров. Данное свойство отличает соответствующие методы кластеризации от метода FCM и позволяет ослаблять влияние аномальных наблюдений на результаты кластеризации.

Заключение

Проведенный анализ целевых функционалов известных методов кластеризации FCM, NC, PCM и FRC раскрывает некоторые неочевидные их свойства и особенности и позволяет сформировать представление об устойчивости решений задач кластеризации по отношению к включению аномальных наблюдений в исследуемую совокупность.

Для анализа проблемы устойчивости решений задачи нечеткой кластеризации в условиях возможного наличия аномальных наблюдений в исследуемой совокупности построены показательные функции относительного влияния произвольного наблюдения, в частности аномального, на оценку разбиения исследуемой совокупности и изучены их свойства. Исследование свойств показательных функций проведено на примерах задач кластеризации с одним и двумя кластерами. Задача с одним основным кластером может послужить основой для определения свойств показательных функций в условиях хорошо разделенных кластеров, находящихся на значительном удалении друг от друга. Задача с двумя основными кластерами описывает свойства показательных функций в области взаимовлияния кластеров. Правомерно воспользоваться полученными результатами и обобщить выявленные свойства показательных функций на случай произвольного числа кластеров.

Основными свойствами показательных функций рассмотренных методов кластеризации является наличие локального экстремума в окрестности каждого из центроидов основных кластеров и монотонное возрастание значений функции при удалении наблюдения от каждого центроида вне области взаимовлияния кластеров. Именно это свойство позволяет находить оптимальное разбиение исследуемой совокупности в условиях отсутствия аномальных наблюдений.

Из отличительных свойств показательных функций методов PCM и FRC следует отметить невыполнение условий экстремума в точках центроидов основных кластеров. Смещение локальных экстремумов относительно соответствующих центроидов незначительно мало только в условиях достаточной разделимости кластеров. Такое смещение заведомо обусловливает несовпадение искомого оптимального расположения центроидов относительно центров тяжести основных кластеров исследуемой совокупности.

Важной особенностью целевых функционалов методов NC, PCM и FRC является ограниченность их показательных функций во всем признаковом пространстве. Это свойство отличает упомянутые целевые функционалы от функционала классического метода нечеткой кластеризации FCM и позволяет ослабить влияние аномальных наблюдений на результаты кластеризации.

Проведенный анализ может помочь в определении исходных значений параметров оптимизационных функционалов для классификации произвольной совокупности. Выявленные характерные свойства показательных функций позволяют построить модифицированные целевые функционалы, учитывающие особенности исследуемой совокупности и ослабляющие влияние аномальных наблюдений. Более того, предложенные в работе показательные функции удобны при сравнении известных функционалов и могут способствовать осмысленному выбору исследователем наиболее соответствующего метода кластеризации при решении практических задач.

Список литературы

1. Bezdek, J.C. Pattern Recognition with Fuzzy Objective Function Algorithms / J.C. Bezdek. – N.Y. : Plenum Press, 1981. – 230 p.
2. Davé, R.N. Robust Clustering Methods: A Unified View / R.N. Davé, R. Krishnapuram // IEEE Transactions on Fuzzy Systems. – 1997. – № 5. – P. 270–293.
3. Dave, R.N. Characterization and detection of noise in clustering / R.N. Dave // Pattern Recognition. – 1991. – Vol. 11, № 12. – P. 657–664.
4. Krishnapuram, R. A possibilistic approach to clustering / R. Krishnapuram, J.M. Keller // IEEE Trans. Fuzzy Systems. – 1993. – № 1. – P. 98–110.
5. Yang, T.-N. Competitive algorithm for the clustering of noisy data / T.-N. Yang, S.-D. Wang // Fuzzy Sets and Systems. – 2004. – № 141. – P. 281–299.
6. Leski, J. Robust Possibilistic Clustering / J. Leski // Archives of control sciences. – 2000. – № 10. – P. 141–155.
7. Rehm, F. A novel approach to noise clustering for outlier detection / F. Rehm, F. Klawonn, R. Kruse // Soft Comput. – 2007. – № 11. – P. 489–494.

Поступила 08.05.09

НПРУП «Агат-Систем»,
Минск, Ф. Скорины, 51
e-mail: kristina.sadovskaya@gmail.com

K. Sadouskaya

ANALYSIS OF THE ROBUSTNESS OF FUZZY CLUSTERING METHODS WITH RESPECT TO ATYPICAL CASES

The paper discusses a problem of fuzzy clustering under conditions of presence of outliers in the input data set. Well-known FCM, NC, PCM, FRC fuzzy clustering optimization methods are analyzed and compared.