УДК 519.8

Л.Ф. ГУЛЯНИЦКИЙ, И.И. РЯСНАЯ

МЕТОД АВТОМАТИЧЕСКОЙ КЛАССИФИКАЦИИ НА БАЗЕ НЕЧЕТКОГО ОТНОШЕНИЯ СХОДСТВА

Аннотация. Для решения задачи автоматической классификации предлагается IFC-метод нечеткой кластеризации, в котором используются новые нечеткие логические операторы — пороговые треугольные нормы и конормы. Данный метод отличается от методов кластеризации на основе нечеткого отношения эквивалентности тем, что позволяет разрабатывать более быстрые алгоритмы построения кластеров. При этом не искажаются данные о связях между элементами исследуемого множества, что обеспечивает прозрачность интерпретации результатов исследований. Приведены примеры применения метода к некоторым известным задачам.

Ключевые слова: нечеткий кластер, классификация, кластерный анализ.

ВВЕДЕНИЕ

После публикации фундаментальных работ Л. Заде [1], С. Тамуры и др. [2] при решении задач автоматической классификации часто используются методы нечеткой кластеризации, составной частью которых является операция транзитивного замыкания матрицы сходства рассматриваемого множества объектов. Результат этой операции — нечеткое отношение эквивалентности. Достоинство такого подхода состоит в том, что любое задаваемое порогом множество этого отношения представляет собой совокупность непересекающихся подмножеств или классов эквивалентности. Упорядоченное множество порогов позволяет определять, как влияет величина порога на количество и состав образуемых классов эквивалентности. Такой подход или его модификации используются как в случаях прямой экспертной оценки сходства, когда двойственное отношение несходства не является метрикой (например, в задаче распознавания портретов [2, 3]), так и в тех случаях, когда исходные данные представлены в виде матрицы расстояний между объектами [4].

Однако дополнение нечеткого отношения эквивалентности является ультраметрикой, свойства которой невозможно интерпретировать в некоторых прикладных задачах [5]. Другая проблема, возникающая при использовании в нечетком кластерном анализе операции транзитивного замыкания матрицы сходства для заданных N объектов, связана с большими вычислительными затратами (порядка $O(N^4)$) для вычисления нечеткого отношения эквивалентности.

В настоящей статье для решения задачи автоматической классификации используется IFC-метод (Interpretable Fuzzy Clusterization) нечеткого кластерного анализа, в котором не используется операция транзитивного замыкания. При этом результаты анализа имеют наглядную интерпретацию, а в алгоритме анализа используется существенно меньшее количество операций. В разд. 1 приводится обоснование предлагаемого IFC-метода. В разд. 2 на основе данных из [2, 4] приведены результаты практического применения IFC-метода. Для нечетких множеств символы $\bigcup$ и $\bigcap$ применяются для нечетких операций объединения (max) и пересечения (min) этих множеств. Символ ■ означает конец доказательства леммы или теоремы.

1. IFC-МЕТОД НЕЧЕТКОГО КЛАСТЕРНОГО АНАЛИЗА

Задача разбиения множества на классы, или задача автоматической классификации, неформально формулируется как задача кластеризации следующим образом [6]: сгруппировать точки заданного множества в подмножества (называемые кластерами) так, чтобы подобные (сходные) точки относились к одному и тому же подмножеству, а не подобные — к различным подмножествам. Подобие (сходство) определяется как рефлексивное и симметричное бинарное отношение на X. В дальнейшем для однозначности такое бинарное отношение будем называть отношением сходства. Э. Руспини показал, что при таком подходе делается попытка установления изоморфизма неизоморфных структур, и предложил новый подход к решению задачи кластеризации [6], сущность которого состоит в аксиоматическом определении исходного понятия «нечеткий кластер» на основе нечеткого отношения сходства, а также разделении процедур кластеризации и процедур формирования исходных нечетких кластеров. Кроме того, в [6] на отношение нечеткого сходства налагаются ограничения, эквивалентные случаю, когда двойственное отношение несходства является метрикой либо псевдометрикой. Задача кластеризации решается на основе такого определения нечеткого кластера, а также применения дополнительных критериев к результирующей системе нечетких подмножеств, представляющих собой результат кластеризации. В данной работе в качестве дополнительного критерия используется связность исходных нечетких кластеров.

Пусть $\tau = \tau(x, y)$ — нечеткое отношение сходства на конечном множестве X, т.е. рефлексивное и симметричное нечеткое бинарное отношение. Пусть также $\hat{\tau}$ — матрица, задающая нечеткое отношение τ. Дополнение нечеткого отношения сходства $\rho(x, y) = 1 - \tau(x, y)$; $x, y \in X$, называется нечетким отношением несходства. Это отношение симметрично и антирефлексивно.

Определение 1. C_x-кластером назовем нечеткое подмножество C_x множества X, $x \in X$, функция принадлежности которого определяется следующим образом: $c_x(y) = \tau(x, y)$, $y \in X$.

В матричном представлении отношения сходства значения $c_x(y)$ — строка матрицы сходства, отражающая сходство элемента $x \in X$ со всеми элементами $y \in X$.

Определение 2. Нечетким отношением сходства уровня α, $\alpha \in (0, 1]$, назовем нечеткое отношение сходства $\tau^{(\alpha)}$, функция принадлежности которого $\forall x, y \in X$ удовлетворяет следующим условиям:

$$\tau^{(\alpha)}(x, y) = \begin{cases} \tau(x, y), & \text{если } \tau(x, y) \geq \alpha, \\ 0, & \text{если } \tau(x, y) < \alpha. \end{cases} \tag{1}$$

Нетрудно видеть, что $\tau^{(\alpha)}(x, y) \leq \tau(x, y)$, т.е. $\tau^{(\alpha)} \subseteq \tau$.

Определение 3. $C_x^{(\alpha)}$-кластером назовем нечеткое подмножество $C_x^{(\alpha)}$ множества X, $x \in X$, $\alpha \in (0, 1]$, функция принадлежности которого определяется следующим образом: $c_x^{(\alpha)}(y) = \tau^{(\alpha)}(x, y)$, $y \in X$.

Определение 4. Нечетким отношением несходства уровня δ, $\delta \in [0, 1)$, назовем нечеткое отношение несходства $\rho^{(\delta)}$, функция принадлежности которого $\forall x, y \in X$ удовлетворяет условиям

$$\rho^{(\delta)}(x, y) = \begin{cases} \rho(x, y), & \text{если } \rho(x, y) \leq \delta, \\ 1, & \text{если } \rho(x, y) > \delta. \end{cases} \tag{2}$$

Очевидно, что $\rho^{(\delta)}(x, y) \geq \rho(x, y)$, т.е. $\rho^{(\delta)} \supseteq \rho$. Если $\delta = 1 - \alpha$, то согласно (1), (2) следует, что $\tau^{(\alpha)}(x, y) + \rho^{(\delta)}(x, y) = 1$.

Треугольной нормой называется функция $T : [0, 1] \times [0, 1] \to [0, 1]$, удовлетворяющая условиям монотонности, ассоциативности и коммутативности, а также равенствам $T(0, 0) = 0$, $T(1, u) = u$ [7]. Треугольной конормой называется функция $S : [0, 1] \times [0, 1] \to [0, 1]$, удовлетворяющая условиям монотонности, ассоциативности и коммутативности, а также равенствам $S(1, 1) = 1$, $S(0, u) = u$ [7].

Для двойственных треугольных норм и конорм имеет место равенство [7]

$$T(u, v) + S(1-u, 1-v) = 1. \tag{3}$$

Определение 5. Пороговой треугольной нормой или треугольной нормой с порогом α, $\alpha \in (0, 1]$, назовем функцию

$$T^{(\alpha)}(u, v) = \begin{cases} T(u, v), & \text{если} \quad T(u, v) \geq \alpha, \\ 0, & \text{если} \quad T(u, v) < \alpha, \end{cases} \tag{4}$$

где $T(u, v)$ — треугольная норма.

Функция $T^{(\alpha)}$ удовлетворяет условиям монотонности, ассоциативности и коммутативности, а также равенству $T^{(\alpha)}(0, 0) = 0$; если $u \geq \alpha$, то $T^{(\alpha)}(1, u) = u$, иначе $T^{(\alpha)}(1, u) = 0$. Кроме того, $T^{(\alpha)}(u, v) \leq T(u, v)$; если $\alpha_1 > \alpha_2$, то $T^{(\alpha_1)}(u, v) \leq T^{(\alpha_2)}(u, v)$.

Определение 6. Пороговой треугольной конормой или треугольной конормой с порогом δ, $\delta \in [0, 1)$, назовем функцию

$$S^{(\delta)}(u, v) = \begin{cases} S(u, v), & \text{если} \quad S(u, v) \leq \delta, \\ 1, & \text{если} \quad S(u, v) > \delta, \end{cases} \tag{5}$$

где $S(u, v)$ — треугольная конорма.

Функция $S^{(\delta)}$ удовлетворяет условиям монотонности, ассоциативности и коммутативности, причем $S^{(\delta)}(1, 1) = 1$; если $u \leq \delta$, то $S^{(\delta)}(0, u) = u$, иначе $S^{(\delta)}(0, u) = 1$. Очевидно, что $S(u, v) \leq S^{(\delta)}(u, v)$. Если $\delta_1 > \delta_2$, то $S^{(\delta_1)}(u, v) \leq S^{(\delta_2)}(u, v)$.

Нетрудно показать, что если T и S — двойственные треугольные норма и конорма соответственно и $\delta = 1 - \alpha$, то пороговые треугольные норма (4) и конорма (5) также двойственные: $T^{(\alpha)}(u, v) + S^{(\delta)}(1-u, 1-v) = 1$.

В дальнейшем полагаем, что $\delta = 1 - \alpha$.

Для конормы Лукасевича [7] $S_L(u, v) = \min(1, u+v)$ пороговая треугольная конорма (5) вычисляется следующим образом:

$$S_L^{(\delta)}(u, v) = \begin{cases} u+v & \text{если} \quad u+v \leq \delta, \\ 1, & \text{если} \quad u+v > \delta. \end{cases} \tag{6}$$

Определение 7. Функцию принадлежности

$$\rho(x, y) = \min_z S_L(\rho(x, z), \rho(z, y)), \; x, y, z \in X,$$

нечеткого отношения несходства будем называть S_L-метрикой на X.

Иначе говоря,

$$S_L(\rho(x, z), \rho(z, y)) \geq S_L(\rho(x, y), \rho(y, y)) = \rho(x, y). \tag{7}$$

Симметричность и антирефлексивность функции принадлежности следует из определения отношения несходства $\rho(x, y)$. Обозначим функцию $S_L(u, v)$ как $u \oplus v$, тогда из (7) получим аналог неравенства треугольника

$$\rho(x, z) \oplus \rho(z, y) \geq \rho(x, y). \tag{8}$$

Теорема 1. Пусть $\rho(x, y)$ — S_L-метрика, тогда $\rho^{(\delta)}(x, y)$ — метрика по пороговой треугольной конорме $S_L^{(\delta)}$.

Доказательство. Следует доказать, что если выполняется условие (8), то

$$S_L^{(\delta)}(\rho^{(\delta)}(x, z), \rho^{(\delta)}(z, y)) \geq S_L^{(\delta)}(\rho^{(\delta)}(x, y), \rho^{(\delta)}(y, y)) = \rho^{(\delta)}(x, y). \tag{9}$$

Рассмотрим возможные случаи.

1. Если $\rho(x, z)+\rho(z, y)>\delta$, то поскольку $\rho(x, y) \leq \rho^{(\delta)}(x, y)$, согласно (6) $S_L^{(\delta)}(\rho^{(\delta)}(x, z), \rho^{(\delta)}(z, y))=1$. Если $\rho(x, y) \leq \delta$, то $\rho^{(\delta)}(x, y)=\rho(x, y)$ и условие (9) выполняется. Если $\rho(x, y)>\delta$, то $\rho^{(\delta)}(x, y)=1$ и условие (9) также выполняется.

2. Если $\rho(x, z)+\rho(z, y) \leq \delta$, то $\rho(x, z) \leq \delta$ и $\rho(z, y) \leq \delta$; согласно (8) имеем $\rho(x, y) \leq \delta$. Из (2) следует, что $\rho^{(\delta)}(x, z)=\rho(x, z)$, $\rho^{(\delta)}(z, y)=\rho(z, y)$, $\rho^{(\delta)}(x, y)=\rho(x, y)$, тогда $S_L^{(\delta)}(\rho^{(\delta)}(x, z), \rho^{(\delta)}(z, y))=S_L(\rho(x, z), \rho(z, y)) \geq$ $\geq \rho(x, y)=\rho^{(\delta)}(x, y)$. ■

Метрику, определяемую по пороговой треугольной конорме $S_L^{(\delta)}$, назовем $S_L^{(\delta)}$-метрикой.

Далее полагаем, что $\rho(x, y)$ — S_L-метрика. Тогда согласно теореме 1 следует, что $\rho^{(\delta)}(x, y)$ является $S_L^{(\delta)}$-метрикой.

Определение 8. Расстоянием на множестве нечетких кластеров $\mathbf{C}=\{C_x | x \in X\}$ назовем функцию $d: \mathbf{C} \times \mathbf{C} \rightarrow [0,1]$ такую, что

$$d(C_x, C_y)=\min_z S_L(1-c_x(z), 1-c_y(z)), \quad x, y, z \in X, \tag{10}$$

где $S_L(u, v)$ — конорма Лукасевича.

Очевидно, что функция d — симметрична и антирефлексивна. Из (10) получаем

$$S_L(1-c_x(z), 1-c_y(z)) \geq S_L(1-c_x(y), 1-c_y(y))=1-c_x(y)=\rho(x, y),$$

т.е. аналог неравенства треугольника. Таким образом, расстояние между кластерами определено как расстояние между ядрами этих кластеров: $d(C_x, C_y)=\rho(x, y)$.

Пусть $\xi(x)=C_x$, $\xi^{-1}(C_x)=x$, $x \in X$, $C_x \in \mathbf{C}$. Очевидно, что биекция ξ между метрическим пространством (X, ρ) и метрическим пространством $(\mathbf{C}, d)$ является изометрией. Обозначим $\mathbf{C}^{(\alpha)}=\{C_x^{(\alpha)} | x \in X\}$. Определим биективное отображение $\xi_\alpha: X \rightarrow \mathbf{C}^{(\alpha)}$ следующим образом: $\xi_\alpha(x)=C_x^{(\alpha)}$ $\forall x \in X$ и обратное отображение $\xi_\alpha^{-1}(C_x^{(\alpha)})=x$ $\forall C_x^{(\alpha)} \in \mathbf{C}^{(\alpha)}$.

Положим $d(C_x^{(\alpha)}, C_y^{(\alpha)})=\rho^{(\delta)}(x, y)$, $C_x^{(\alpha)}, C_y^{(\alpha)} \in \mathbf{C}^{(\alpha)}$, где $\rho^{(\delta)}(x, y)$ — $S_L^{(\delta)}$-метрика. Тогда биекция ξ_α между метрическим пространством $(X, \rho^{(\delta)})$ и метрическим пространством $(\mathbf{C}^{(\alpha)}, d)$ является изометрией. Таким образом, на основе введения новых нечетких логических операторов — пороговых треугольных норм и конорм — на множестве $\mathbf{C}^{(\alpha)}$ определена метрика, т.е. расстояние между двумя произвольными кластерами $C_x^{(\alpha)}, C_y^{(\alpha)} \in \mathbf{C}^{(\alpha)}$ равно расстоянию $\rho^{(\delta)}(x, y)$ между ядрами этих кластеров. Задание метрики на множестве кластеров $\mathbf{C}^{(\alpha)}$ дает новые возможности для исследования структуры множества X.

Определение 9. Кластеры $C_x^{(\alpha)}, C_y^{(\alpha)} \in \mathbf{C}^{(\alpha)}$, $x, y \in X$, назовем соседними, если $y \in \operatorname{supp} C_x^{(\alpha)}$, $x \in \operatorname{supp} C_y^{(\alpha)}$.

Для соседних кластеров согласно (1), (2) $\tau^{(\alpha)}(x, y)=\tau(x, y) \geq \alpha$ и $\rho^{(\delta)}(x, y)=$ $=\rho(x, y) \leq \delta$. Следовательно, $d(C_x^{(\alpha)}, C_y^{(\alpha)})=\rho^{(\delta)}(x, y) \leq \delta$. Если кластеры $C_x^{(\alpha)}$, $C_y^{(\alpha)}$ не являются соседними, то $\tau^{(\alpha)}(x, y)=0$, $\rho^{(\delta)}(x, y)=1$, $d(C_x^{(\alpha)}, C_y^{(\alpha)})=1$.

Определение 10. δ-окрестностью кластера $C_x^{(\alpha)}$ назовем множество кластеров $\mathbf{O}_x=\{C_y^{(\alpha)} | y \in \operatorname{supp} C_x^{(\alpha)}, y \neq x\}$.

Если $\mathbf{O}_x \neq \varnothing$, то $\forall C_y^{(\alpha)} \in \mathbf{O}_x$ имеем $\rho^{(\delta)}(x, y) \leq \delta$. Следовательно, любой из этих кластеров является соседним с кластером $C_x^{(\alpha)}$.

Определение 11. Кластеры $C_x^{(\alpha)}, C_y^{(\alpha)} \in \mathbf{C}^{(\alpha)}$, $x, y \in X$, назовем смежными, если $\bigcap (C_x^{(\alpha)}, C_y^{(\alpha)}) \neq \varnothing$, а $\rho^{(\delta)}(x, y) = 1$.

Для смежных кластеров существует хотя бы один общий соседний кластер $C_z^{(\alpha)}$ такой, что $z \in \bigcap (\operatorname{supp} C_x^{(\alpha)}, \operatorname{supp} C_y^{(\alpha)})$. Если $\bigcap (C_x^{(\alpha)}, C_y^{(\alpha)}) \neq \varnothing$, то эти кластеры либо соседние, либо смежные.

Определение 12. Кластеры $C_x^{(\alpha)}, C_y^{(\alpha)} \in \mathbf{C}^{(\alpha)}$ назовем связанными, если существует такая совокупность кластеров $C_x^{(\alpha)} = C_{z_0}^{(\alpha)}, C_{z_1}^{(\alpha)}, C_{z_2}^{(\alpha)}, \ldots, C_{z_{n-1}}^{(\alpha)}, C_{z_n}^{(\alpha)} = C_y^{(\alpha)}$, что $\bigcap (C_{z_0}^{(\alpha)}, C_{z_1}^{(\alpha)}) \neq \varnothing$, $\bigcap (C_{z_1}^{(\alpha)}, C_{z_2}^{(\alpha)}) \neq \varnothing, \ldots, \bigcap (C_{z_{n-1}}^{(\alpha)}, C_{z_n}^{(\alpha)}) \neq \varnothing$; $C_{z_i}^{(\alpha)} \in \mathbf{C}^{(\alpha)}$, $i \in \{0, \ldots, n\}$.

Очевидно, что соседние и смежные кластеры являются связанными. Однако связанные кластеры не являются в общем случае соседними или смежными.

Теорема 2. Между связанными кластерами $C_x^{(\alpha)}, C_y^{(\alpha)} \in \mathbf{C}^{(\alpha)}$ существует путь, проходящий через соседние кластеры.

Поскольку для любых соседних кластеров $C_x^{(\alpha)}, C_y^{(\alpha)} \in \mathbf{C}^{(\alpha)}$ имеет место $d(C_x^{(\alpha)}, C_y^{(\alpha)}) = \rho^{(\delta)}(x, y) \leq \delta$, то назовем такой путь δ-связным.

Доказательство. Пусть $n = 1$ (см. определение 12), тогда $\bigcap (C_x^{(\alpha)}, C_y^{(\alpha)}) \neq \varnothing$. Кластеры $C_x^{(\alpha)}$, $C_y^{(\alpha)}$ могут быть соседними или смежными. В первом случае $\rho^{(\delta)}(x, y) \leq \delta$. Во втором случае существует общий соседний кластер $C_z^{(\alpha)}$ такой, что $C_z^{(\alpha)} \in \mathbf{O}_x$ и $C_z^{(\alpha)} \in \mathbf{O}_y$, $z \in \operatorname{supp} C_x^{(\alpha)}$, $z \in \operatorname{supp} C_y^{(\alpha)}$. Иначе говоря, $\rho^{(\delta)}(x, z) \leq \delta$, $\rho^{(\delta)}(z, y) \leq \delta$ и путь между кластерами $C_x^{(\alpha)}, C_y^{(\alpha)}$ — δ-связный.

Пусть кластеры $C_x^{(\alpha)} = C_{z_0}^{(\alpha)}$ и $C_{z_{k-1}}^{(\alpha)}$ связаны, $C_{z_k}^{(\alpha)} = C_y^{(\alpha)}$, $k < n$, тогда аналогично доказывается существование δ-связного пути между кластерами $C_{z_{k-1}}^{(\alpha)}$ и $C_{z_k}^{(\alpha)}$, т.е. кластеры $C_x^{(\alpha)}, C_y^{(\alpha)}$ связаны. По индукции следует существование δ-связного пути для $k = n$. ■

Если между двумя кластерами не существует δ-связного пути, то эти кластеры назовем несвязанными.

Определение 13. Связным предклассом назовем множество кластеров $\{C_x^{(\alpha)}\}$, в котором любые два кластера связаны.

Другими словами, между любыми двумя элементами связного предкласса существует δ-связный путь. Очевидно, что множество кластеров $\mathbf{O}_x$ представляет связный предкласс.

Определение 14. Связным классом или Δ-кластером назовем максимальный предкласс.

Обозначим мощность $|X| = N$.

Теорема 3. Множество Δ-кластеров $\Sigma = \{\Delta_j\}_{j=1}^m$, $2 \leq m \leq N$, представляет собой разбиение множества $\mathbf{C}^{(\alpha)}$.

Доказательство. Предположим, что Δ-кластеры пересекаются, т.е. $\bigcap (\Delta_i, \Delta_j) \neq \varnothing$, $i \neq j$, $i, j \in \{1, \ldots, m\}$. Тогда существует кластер $C_z^{(\alpha)} \in \bigcap (\Delta_i, \Delta_j)$ и кластеры Δ_i, Δ_j оказываются связанными, т.е. не являются максимальными предклассами, что противоречит определению 14. Обозначим $|\Delta_j| = k_j$, $j = \overline{1, m}$.

Если $k_j = 1$, то кластер Δ_j содержит только один элемент. Так как $\forall C_x^{(\alpha)} \in \mathbf{C}^{(\alpha)}$ $\exists \Delta_j \in \Sigma$ и $\forall \Delta_j \in \Sigma \, \exists C_x^{(\alpha)} \in \mathbf{C}^{(\alpha)}$, то $\bigcup_{j=1}^{m} \Delta_j = \mathbf{C}^{(\alpha)}$ и $\sum_{j=1}^{m} k_j = N$. ■

Очевидно, что при фиксированном пороге $\alpha = 1 - \delta$ построение множества Δ-кластеров $\Sigma = \{\Delta_i\}_{i=1}^{m}$, $2 \le m \le N$, представляет решение задачи автоматической классификации множества объектов X. Характеристическое свойство Δ-кластеров состоит в наличии δ-связного пути между любыми элементами одного и того же кластера и отсутствии такого пути между элементами различных кластеров.

Опишем один из возможных вариантов алгоритма IFC-метода нечеткой кластеризации при фиксированной системе порогов. Пусть теперь $X = \{x_i\}_{i=1}^{N}$ — множество объектов и задано нечеткое отношение τ, представленное матрицей сходства $\hat{\tau} = (\tau(x_i, x_j))_{i=1,\, j=1}^{N,\, N}$. Находим пару объектов с минимальным значением сходства $\alpha_{\min} > 0$ и вычисляем значения порогов $\alpha = \{\alpha_k\}_{k=1}^{p}$; $\alpha_1 > \ldots > \alpha_k > \ldots > \alpha_p$; $\alpha_k = \alpha_{\min} + k(1 - \alpha_{\min}) / p$, $k = 1, 2, \ldots, p$. Представим алгоритм анализа кластерной структуры множества $X = \{x_i\}_{i=1}^{N}$ для фиксированного значения $\alpha = \alpha_k$ и матрицы сходства $\hat{\tau}$.

1. Для значения порога α_k строится множество кластеров $\{C_{x_i}^{(\alpha_k)}\}_{i=1}^{N}$.

2. Вычисляется совокупность мощностей носителей каждого кластера.

3. Определяется кластер с максимальной мощностью носителя. Если таких кластеров несколько, то берется любой из них. Далее на основе объединения носителей связанных кластеров строится множество Δ-кластеров.

Отметим, что на этапах 2 и 3 можно вычислять мощность кластеров $\{C_{x_i}^{(\alpha_k)}\}_{i=1}^{N}$, Δ-кластеров и их носителей. Эта информация может быть полезной при содержательной интерпретации результатов, а также для приближенного решения задачи автоматической классификации при большом объеме исходных данных в конкретных приложениях. Вычислительная сложность алгоритма не превышает $O(N^2)$.

2. ПРИМЕРЫ ПРАКТИЧЕСКОГО ПРИМЕНЕНИЯ

На рис. 1 приведены кластеры, полученные в результате применения IFC-метода для набора точек на плоскости (см. [4, табл. 7.3]) при выборе $\alpha = 0{,}6$. Функция принадлежности нечеткого отношения сходства τ для точек x_i, x_j вычисляется на основе нормировки евклидового расстояния $d(x_i, x_j)$ делением на максимальное значение расстояния

$$\tau(x_i, x_j) = 1 - d(x_i, x_j) / \max_{i, j \in \overline{1, N}} d(x_i, x_j) . \qquad (11)$$

Этот набор данных аппроксимируется кластерами эллипсовидной формы, которые могут быть получены как известным методом Fuzzy C-Means, так и методом, содержащим операцию транзитивного замыкания [4].

На рис. 2 приведен набор данных, состоящий из 40 точек, размещенных на двух концентрических окружностях [4, табл. 7.12]. Метод Fuzzy C-Means не достигает результата для такого набора данных, поэтому в [4] для получения разбиения, состоящего из точек, лежащих на одной и той же окружности, применяется операция транзитивного замыкания. Используем функцию принадлежности нечеткого отношения сходства (11) на базе IFC-метода и получаем такое разбиение для значения $\alpha = 0{,}8$.

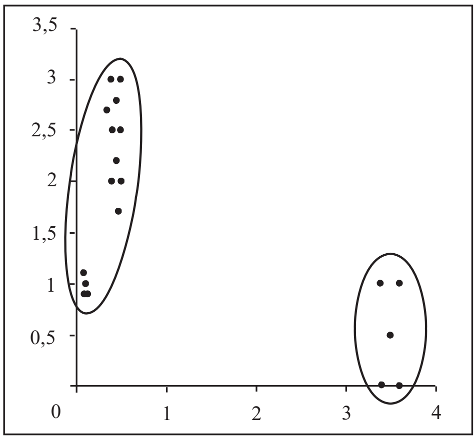

Рис. 1

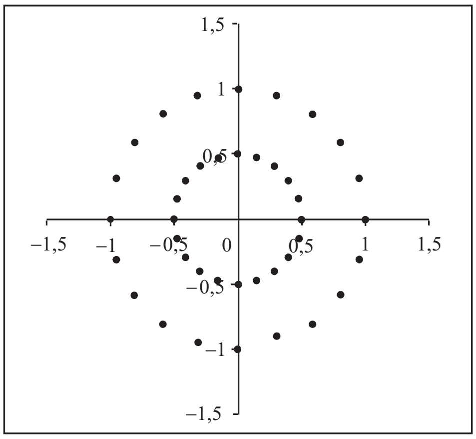

Рис. 2

Таблица 1

	Ch1_A	Ch1_B	Ch1_C	Gm_C	P1_B	P1_A	Ch2_B	Ch2_A	Ch2_C	P1_C	P2_B	Ch3_C	P2_A	Ch3_B	P2_C	Ch3_A
Ch1_A	1															
Ch1_B	0	1														
Ch1_C	0	0	1													
Gm_C	0	0	0,4	1												
P1_B	0	0,8	0	0	1											
P1_A	0,5	0	0,2	0,2	0	1										
Ch2_B	0	0,8	0	0	0,4	0	1									
Ch2_A	0,4	0,2	0,2	0,5	0	0,8	0	1								
Ch2_C	0	0,4	0	0,8	0,4	0,2	0,4	0	1							
P1_C	0	0	0,2	0,2	0	0	0,2	0	0,2	1						
P2_B	0	0,5	0,2	0,2	0	0	0,8	0	0,4	0,2	1					
Ch3_C	0	0	0,2	0,8	0	0	0	0	0,4	0,8	0	1				
P2_A	0,8	0	0,2	0,4	0	0,4	0	0,4	0	0	0	0	1			
Ch3_B	0	0,8	0	0,2	0,4	0	0,8	0	0,2	0,2	0,6	0	0	1		
P2_C	0	0	0,4	0,8	0	0,2	0	0	0,2	0	0	0,2	0,2	0	1	
Ch3_A	0,6	0	0	0,2	0,4	0,8	0	0,4	0	0	0	0	0,4	0,2	0	1

Представим результаты использования IFC-метода в задаче распознавания портретов, принадлежащих одной и той же семье. Отметим, что эта задача существенным образом отличается от сформулированной выше неформальной постановки задачи кластеризации. Согласно неформальному определению кластеры должны содержать только сходные элементы, в то время как классы эквивалентности в задаче распознавания портретов кроме сходных содержат и несходные элементы (портреты), поскольку члены семьи могут не иметь сходства между собой. Данные по сходству 16 портретов, приведенные в табл. 1, взяты из работы [2]. Семья A состоит из двух родителей (Parent) и трех детей (Child): P1_A, P2_A, Ch1_A, Ch2_A, Ch3_A; семья B имеет следующих представителей: P1_B, P2_B, Ch1_B, Ch2_B, Ch3_B; в семью C входит еще и бабушка (Grandmother): P1_C, P2_C, Ch1_C, Ch2_C, Ch3_C, Gm_C.

В табл. 2 приведены результаты обработки данных для $\alpha = 0{,}6$. Прямоугольником выделены полученные данные для каждой отдельной семьи. Согласно табл. 2 семьи A и B распознаются безошибочно, однако в семью C не включен один ребенок (Ch1_C), поскольку для данного порога этот элемент матрицы оказывается изолированным, т.е. не сходным с другими элементами матрицы. Такой результат классификации полностью совпадает с результатами в [2], однако при этом не используется операция транзитивного замыкания.

Поскольку IFC-метод связи между элементами не изменяет, то можно проводить дальнейшие исследования. В частности, при снижении порога нечеткий

Таблица 2

	P1_A	P2_A	Ch_1A	Ch2_A	Ch3_A	P1_B	P2_B	Ch1_B	Ch2_B	Ch3_B	Gm_C	P1_C	P2_C	Ch1_C	Ch2_C	Ch3_C
P1_A	1	0	0	0,8	0,8	0	0	0	0	0	0	0	0	0	0	0
P2_A	0	1	0,8	0	0	0	0	0	0	0	0	0	0	0	0	0
Ch1_A	0	0,8	1	0	0,6	0	0	0	0	0	0	0	0	0	0	0
Ch2_A	0,8	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
Ch3_A	0,8	0	0,6	0	1	0	0	0	0	0	0	0	0	0	0	0
P1_B	0	0	0	0	0	1	0	0,8	0	0	0	0	0	0	0	0
P2_B	0	0	0	0	0	0	1	0	0,8	0,6	0	0	0	0	0	0
Ch1_B	0	0	0	0	0	0,8	0	1	0,8	0,8	0	0	0	0	0	0
Ch2_B	0	0	0	0	0	0	0,8	0,8	1	0,8	0	0	0	0	0	0
Ch3_B	0	0	0	0	0	0	0,6	0,8	0,8	1	0	0	0	0	0	0
Gm_C	0	0	0	0	0	0	0	0	0	0	1	0	0,8	0 (*0,4*)	0,8	0,8
P1_C	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0,8
P2_C	0	0	0	0	0	0	0	0	0	0	0,8	0	1	0 (*0,4*)	0	0
Ch1_C	0	0	0	0	0	0	0	0	0	0	0 (*0,4*)	0	0 (*0,4*)	1	0	0
Ch2_C	0	0	0	0	0	0	0	0	0	0	0,8	0	0	0	1	0
Ch3_C	0	0	0	0	0	0	0	0	0	0	0,8	0	0	0	0	1

кластер $C_{\text{Ch1\_C}}^{(0,4)}$ оказывается связанным с семьей C (курсивные цифры в табл. 2). Значит, состав всех семей определяется правильно. Такую дополнительную процедуру поиска «ближайшего» класса эквивалентности можно использовать для классификации изолированных элементов. Результат классификации оказывается более адекватным, чем при использовании операции $\max-\min$ транзитивного замыкания [2], а полученная классификация легко интерпретируется.

ЗАКЛЮЧЕНИЕ

Решение и интерпретация результатов предложенного подхода к автоматической классификации на основе IFC-метода базируется на введении нечетких логических операторов — пороговых треугольных норм и конорм. Данный метод отличается от методов, основанных на нечетком отношении эквивалентности, поскольку позволяет разрабатывать более быстрые алгоритмы построения классов, так как не используется операция транзитивного замыкания. Кроме того, не искажаются данные о связях между элементами исходного множества, что обеспечивает прозрачность интерпретации получаемых результатов и позволяет уточнять их при дальнейших исследованиях структуры классов.

Возможное направление дальнейших исследований: вопросы адекватности результатов автоматической классификации в тех случаях, когда нечеткое отношение сходства задается на основе характеристик классифицируемых объектов, измеренных в различных количественных и качественных шкалах.

СПИСОК ЛИТЕРАТУРЫ

1. Zadeh L.A. Similarity relations and fuzzy ordering // Information Sciences. — 1971. — **3.** — P. 177–200.
2. Tamura S., Higuchi S., Tanaka K. Pattern classification based on fuzzy relations // IEEE Trans. Systems, Man and Cybernetics. — 1971. — **SMC-1**. — P. 61–66.
3. Yang M.-S., Shih H.-M. Cluster analysis based on fuzzy relations // Fuzzy Sets and Systems. — 2001. — **120.** — P. 197–212.
4. Барсегян А.А., Куприянов М.С., Степаненко В.В. и др. Методы и модели анализа данных: OLAP и Data Mining. — СПб.: БХВ-Петербург, 2004. — 336 с.
5. Кофман А. Введение в теорию нечетких множеств. — М.: Радио и связь, 1982. — 432 с.
6. Руспини Э.Г. Последние достижения в нечетком кластер-анализе / Нечеткие множества и теория возможностей. Последние достижения. — М.: Радио и связь, 1986. — С. 114–132.
7. Нечеткие множества в моделях управления и искусственного интеллекта / Под ред. Д.А. Поспелова. — М.: Наука, 1986. — 312 с.

Поступила 06.05.2015