Управляемые системы, ИМ СО АН СССР, 1980 г., вып.20

УДК 519.95.

ОБ ОДНОЙ ЗАДАЧЕ КЛАССИФИКАЦИИ

А.В.Козина

Классификация многомерных объектов – одна из основных задач обработки данных, например социологических. Так, задачей кластерного (классификационного) анализа является разбиение множества объектов, описываемых набором признаков, на K классов (кластеров) таким образом, чтобы "похожие" объекты принадлежали одному и тому же классу, а "непохожие" – разным. Критерием оптимальности может быть функционал, выражающий качество разбиения. Задавая вид этого функционала, который диктуется содержательным анализом, получаем различные задачи классификации (кластеризации) [1, 2].

§ 1. Постановка задачи

Пусть заданы признаки $T_1, \ldots, T_m$ и P_j $(j = 1, \ldots, m)$ – множество значений j -го признака. Обозначим через P^m множество всевозможных наборов $a = (a_1, \ldots, a_m)$ длины m, j -е компоненты которых принадлежат P_j, $j = 1, \ldots, m$, мощность этого множества будет $|P^m| = \prod_{j=1}^{m} |P_j|$. Далее, пусть X – произвольное множество мощности n наборов из P^m, не обязательно различных, элементы которого пронумерованы в некотором порядке, т.е.

$$X = \{x^i = (x_1^i, \ldots, x_m^i) \mid (x_1^i, \ldots, x_m^i) \in P^m,\ i = 1, \ldots, n\}.$$

В дальнейшем элементы множества X будем называть объектами. Тогда для объекта $x^i = (x_1^i, \ldots, x_m^i)$ компоненту x_j^i набора назовем значением признака T_j на объекте x^i. По определению, будем считать, что объекты x^s и x^t различаются на множестве X, если $s \neq t$. Разбиением множества X назовем такое семейство его непустых подмножеств $X_s \subseteq X$, что $X = \bigcup_{s=1}^{k} X_s$, $X_s \cap X_t = \emptyset$ для $s \neq t$, и обозначим это разбиение через $A(X) = \{X_1, \ldots, X_k\}$, подмножества X_s назовем классами. Задача заключается в том, чтобы найти оптимальное в смысле некоторого критерия разбиение множества X объектов на классы. Введем этот критерий.

Определим меру различия на множестве P^m. Для наборов

$$a = (a_1, \ldots, a_m),\ b = (b_1, \ldots, b_m) \in P^m$$

имеем

$$\rho(a,b) = \frac{1}{m}\sum_{j=1}^{m}\varepsilon_j(a_j,b_j),$$

где ε_j - функционал, определенный на парах элементов множества P_j и обладающий свойствами: $\varepsilon_j(a_j,b_j)\geq 0$, $\varepsilon_j(a_j,b_j)=\varepsilon_j(b_j,a_j)$ и $\varepsilon_j(a_j,a_j)=0$. В зависимости от выбранных ε_j функционал $\rho(a,b)$ может быть метрикой (расстоянием), что превращает P^m в метрическое пространство. Конкретный вид функционала ε_j, вообще говоря, определяется содержательной постановкой задачи и в данной работе не рассматривается. Если $x^i=(x_1^i,\dots,x_m^i)\in X$ и $a=(a_1,\dots,a_m)\in P^m$, то, по определению, имеем $\rho(x^i,a)=\frac{1}{m}\sum_{j=1}^{m}\varepsilon_j(x_j^i,a_j)$.

Центром произвольного подмножества X_0 множества X называется набор $C\in P^m$, для которого

$$\sum_{x^i\in X_0}\rho(x^i,c) = \min_{a\in P^m}\sum_{x^i\in X_0}\rho(x^i,a). \qquad /1/$$

Величину

$$\Phi(X_0)=\frac{1}{|X_0|}\sum_{x^i\in X_0}\rho(x^i,c) \qquad /2/$$

назовем степенью неоднородности множества X_0. Степенью неоднородности разбиения $\mathcal{A}(X)=\{X_1,\dots,X_K\}$ называется функционал

$$\Phi(\mathcal{A})=\sum_{s=1}^{K}\frac{|X_s|}{n}\Phi(X_s). \qquad /3/$$

Из /1/, /2/ и /3/ следует, что

$$\Phi(\mathcal{A})=\frac{1}{n}\sum_{s=1}^{K}\sum_{x_i\in X_s}(\rho\ x^i,c^s),$$

где C^s - центр множества X_s, $s=1,\dots,K$. Обозначив через $\mathcal{A}(X)$ множество всевозможных разбиений множества X, назовем разбиение $\mathcal{A}(X)\in\mathcal{A}'(X)\subseteq\mathcal{A}(X)$ оптимальным на множестве $\mathcal{A}'(X)$, если

$$\Phi(\mathcal{A})=\min_{\mathcal{A}'\in\mathcal{A}'(X)}\Phi(\mathcal{A}').$$

Ставится задача: построить разбиение, оптимальное на множестве $\mathcal{A}_K(X)$ всех разбиений на K классов.

§ 2. Некоторые свойства задачи.

Пусть для фиксированного натурального K задана последовательность наборов $C=\{c^1,\dots,c^K\}$, $c^s\in P^m$, $s=1,\dots,K$. Тогда через $\mathcal{A}(X,C)$ обозначим множество всех таких разбиений $\mathcal{A}(X)=\{X_1,\dots,X_K\}$, что для каждого $s=1,\dots,K$ набор c^s является центром класса X_s. Будем говорить, что разбиение $\mathcal{A}(X)=\{X_1,\dots,X_K\}$ порождено после-

довательностью наборов C, если для каждого $s = 1, \dots, K$ и каждого объекта $x^i \in X_s$ выполняется неравенство $\rho(x^i, c^s) \leq \rho(x^i, c^t)$ для всех $t = 1, \dots, K$.

Нетрудно убедиться в том, что для любого разбиения $A_1(X) = \{X_1^1, \dots, X_K^1\} \in \mathcal{A}(X, C)$ и любого разбиения $A_2(X) = \{X_1^2, \dots, X_K^2\}$, порожденного последовательностью наборов C, справедливо неравенство $\Phi(A_2) \leq \Phi(A_1)$, причем если $A_2 \notin \mathcal{A}(X, C)$, то $\Phi(A_2) < \Phi(A_1)$.

Действительно, для произвольной фиксированной последовательности наборов $C = \{c^1, \dots, c^K\}$ положим

$$\Psi(A) = \frac{1}{n} \sum_{s=1}^{K} \sum_{x^i \in X_s} \rho(x^i, c^s),$$

где $A(X) = \{X_1, \dots, X_K\}$. Тогда для разбиения $A(X) \in \mathcal{A}(X, C)$ выполняется равенство $\Psi(A) = \Phi(A)$. Если же $A(X) \notin \mathcal{A}(X, C)$, то найдется такое целое t, $1 \leq t \leq K$, что c^t не является центром класса X_t. Поэтому

$$\sum_{x^i \in X_t} \rho(x^i, \hat{c}^t) < \sum_{x^i \in X_t} \rho(x^i, c^t),$$

$$\sum_{x^i \in X_s} \rho(x^i, \hat{c}^s) \leq \sum_{x^i \in X_s} \rho(x^i, c^s) \quad \text{для всех } s \neq t,$$

где $\hat{c}^s$ – центр класса X_s, $s = 1, \dots, K$.
Следовательно,

$$\Phi(A) = \frac{1}{n} \sum_{s=1}^{K} \sum_{x^i \in X_s} \rho(x^i, \hat{c}^s) < \frac{1}{n} \sum_{s=1}^{K} \sum_{x^i \in X_s} \rho(x^i, c^s) = \Psi(A).$$

Для доказательства нашего утверждения достаточно показать, что для разбиения $A_2(X) = \{X_1^2, \dots, X_K^2\}$, порожденного последовательностью наборов C, верно $\Psi(A_2) \leq \Psi(A)$ для всех $A \in \mathcal{A}_K(X)$.

Действительно, если это так, то найдется разбиение $A(X) = \{X_1, \dots, X_K\}$, для которого верно неравенство

$$\Psi(A_2) = \frac{1}{n} \sum_{s=1}^{K} \sum_{x^i \in X_s^2} \rho(x^i, c^s) > \frac{1}{n} \sum_{s=1}^{K} \sum_{x^i \in X_s} \rho(x^i, c^s) = \Psi(A).$$

Тогда можно отыскать объект $x^i \in X$, принадлежащий в разбиении $A_2(X)$ некоторому классу X_s^2, а в разбиении $A(X)$ – классу X_t, для которого $\rho(x^i, c^s) > \rho(x^i, c^t)$. С другой стороны, так как разбиение $A_2(X)$ порождено последовательностью наборов C, объект $x^i \in X_s^2$ удовлетворяет неравенству $\rho(x^i, c^s) \leq \rho(x^i, c^t)$. Из полученного противоречия следует, что $\Psi(A_2) \leq \Psi(A)$ для всех $A \in \mathcal{A}_K(X)$.

Из приведенных выше рассуждений следует:

I) если разбиение $A(X)$ с центрами классов $C = \{c^1, \dots, c^K\}$ является оптимальным на множестве $\mathcal{A}_K(X)$, то оно порождается последовательностью наборов C, и любое разбиение $A'(X)$, порожденное последовательностью наборов C, оптимально на множестве

$\alpha_K(X)$;

2) справедливо равенство

$$\min_{\mathcal{A} \in \alpha_K(X)} \Phi(\mathcal{A}) = \min_{C} \Phi(\mathcal{A}_C),$$

где C пробегает множество всевозможных последовательностей наборов из множества P^m и $\mathcal{A}_C$ - разбиение, порожденное последовательностью наборов C.

Последнее утверждение весьма полезно для нахождения оптимального разбиения в случае, когда число объектов велико по сравнению с числом признаков, числом значений признаков и числом классов. Например, если признаки - дихотомические, то число всевозможных наборов последовательностей K наборов C равно 2^{Km}, тогда как число всех разбиений $2^{n \log_2 K}$.

Целью дальнейшего рассмотрения является установление связи между решением задачи о построении оптимального разбиения на множестве признаков $T_1, \dots, T_m$ и решением той же задачи на некотором подмножестве этого множества.

Для множеств натуральных чисел $L = \{p_1, \dots, p_\ell \mid p_1 < \dots < p_\ell\}$ и $H = \{q_1, \dots, q_{m-\ell} \mid q_1 < \dots < q_{m-\ell}\}$ таких, что $L \cap H = \emptyset$ и $L \cup H = \{1, \dots, m\}$, определим на множестве P^m отображения φ_L и φ_H следующим образом: для набора $a = (a_1, \dots, a_m) \in P^m$ положим $\varphi_L(a) = (a_{p_1}, \dots, a_{p_\ell})$ и $\varphi_H(a) = (a_{q_1}, \dots, a_{q_{m-\ell}})$. Для объекта $x^i = (x^i_1, \dots, x^i_m) \in X$ положим $\varphi_L(x^i) = (x^i_{p_1}, \dots, x^i_{p_\ell})$. Образ множества X при отображении φ_L назовем L-редукцией множества X. Напомним, что на множестве X объекты x^s и x^t различны, если $s \neq t$. При отображении φ_L объекты сохраняют свои порядковые номера, и если $x^s \neq x^t$, т.е. $s \neq t$, то $\varphi_L(x^s) \neq \varphi_L(x^t)$. Поэтому на множестве X отображение φ_L - взаимно-однозначно. Следовательно, на множестве $\varphi_L(X)$ можно определить отображение φ_L^{-1}: $\varphi_L^{-1}(\tilde{x}^i) = x^i$, если $\tilde{x}^i = \varphi_L(x^i)$. Переход от множества X к множеству $\varphi_L(X)$ означает исключение из рассмотрения признаков $T_{q_1}, \dots, T_{q_{m-\ell}}$.

На множестве $P^\ell = \varphi_L(P^m)$ редуцированных наборов и множестве $\varphi_L(X)$ редуцированных объектов так же, как на множествах P^m и X соответственно, вводятся функционал ρ, понятия центра множества, разбиения и степени неоднородности. Таким образом, для любой редукции можно поставить задачу оптимального разбиения.

Для произвольных наборов $a = (a_1, \dots, a_m)$, $b = (b_1, \dots, b_m) \in P^m$ имеем

$$m\rho(a,b) = \sum_{j=1}^{m} \varepsilon_j(a_j, b_j) = \sum_{j \in L} \varepsilon_j(a_j, b_j) + \sum_{j \in H} \varepsilon_j(a_j, b_j) =$$

$$= \ell \cdot \rho(\varphi_L(a), \varphi_L(b)) + (m - \ell)\rho(\varphi_H(a), \varphi_H(b)).$$

Из этого равенства вытекает следующая

Л е м м а I. Набор C является центром множества $X_0 \subseteq X$ тогда и только тогда, когда $\varphi_L(C)$ – центр множества $\varphi_L(X_0)$ и $\varphi_H(C)$ – центр множества $\varphi_H(X_0)$.

Для разбиения $\mathcal{A}(X) = \{X_1, \dots, X_K\}$ и множества $L \subseteq \{1, \dots, m\}$ через $\varphi_L(\mathcal{A})$ обозначим новое разбиение $\varphi_L(\mathcal{A}) = \{\varphi_L(X_1), \dots, \varphi_L(X_K)\}$ и соответственно $\varphi_L^{-1}(\tilde{\mathcal{A}}) = \{\varphi_L^{-1}(\tilde{X}_1), \dots, \varphi_L^{-1}(\tilde{X}_K)\}$, где $\tilde{\mathcal{A}} = \{\tilde{X}_1, \dots, \tilde{X}_K\}$.

Л е м м а 2. Для произвольного разбиения $\mathcal{A}(X)$ и разбиений $\mathcal{A}_1 = \varphi_L(\mathcal{A})$ и $\mathcal{A}_2 = \varphi_H(\mathcal{A})$ верно равенство $m\,\Phi(\mathcal{A}) = \ell \cdot \Phi(\mathcal{A}_1) + (m-\ell)\Phi(\mathcal{A}_2)$.

Д о к а з а т е л ь с т в о. Пусть классы разбиения $\mathcal{A}(X) = \{X_1, \dots, X_K\}$ имеют центры $C^1, \dots, C^K$. По определению степени неоднородности, имеем

$$m\,\Phi(\mathcal{A}) = \frac{1}{n} \sum_{s=1}^{K} \sum_{x^i \in X_s} m\,\rho(x^i, C^s) =$$

$$= \frac{1}{n} \sum_{s=1}^{K} \sum_{x^i \in X_s} \left[\ell\,\rho\left(\varphi_L(x^i), \varphi_L(C^s)\right) + (m-\ell)\,\rho\left(\varphi_H(x^i), \varphi_H(C^s)\right)\right] =$$

$$= \frac{\ell}{n} \sum_{s=1}^{K} \sum_{\varphi_L(x^i) \in \varphi_L(X_s)} \rho\left(\varphi_L(x^i), \varphi_L(C^s)\right) +$$

$$+ \frac{m-\ell}{n} \sum_{s=1}^{K} \sum_{\varphi_H(x^i) \in \varphi_H(X_s)} \rho\left(\varphi_H(x^i), \varphi_H(C^s)\right) = \ell \cdot \Phi(\mathcal{A}_1) + (m-\ell)\,\Phi(\mathcal{A}_2).$$

Лемма доказана.

Назовем степенью K -неоднородности множества X величину

$$\Phi_K(X) = \min_{\mathcal{A} \in \mathfrak{A}_K(X)} \Phi(\mathcal{A}).$$

Заметим, что степень I-неоднородности – это ранее введенная степень неоднородности множества X :

$$\Phi_1(X) = \Phi(X) = \frac{1}{n} \sum_{x^i \in X} \rho(x^i, C),$$

где C – центр множества X.

Справедливы следующие леммы.

Л е м м а 3. Если набор $C \in P^m$ является центром каждого класса X_s, $s = 1, \dots, K$ разбиения $\mathcal{A}(X) = \{X_1, \dots, X_K\}$, то C – центр множества X и $\Phi(\mathcal{A}) = \Phi(X)$.

Л е м м а 4. Для любого натурального K и любого разбиения $\mathcal{A} \in \mathfrak{A}_K(X)$ верно неравенство $\Phi_K(X) \leq \Phi(\mathcal{A}) \leq \Phi(X)$.

Т е о р е м а I. Для любого натурального K и любого множества X объектов верно неравенство

$$(m-\ell)\,\Phi_K(\varphi_H(X)) \leq m\,\Phi_K(X) - \ell\,\Phi_K(\varphi_L(X)) \leq (m-\ell)\,\Phi(\varphi_H(X)).$$

Доказательство. I. Пусть разбиение $\mathcal{A}(X)$ оптимально на множестве $\mathcal{A}(X)$ и $\mathcal{A}_1 = \varphi_L(A)$, $\mathcal{A}_2 = \varphi_H(\mathcal{A})$. Тогда из лемм 2 и 4 получаем

$$m\,\Phi_K(X) = m\,\Phi(\mathcal{A}) = l\,\Phi(\mathcal{A}_1) + (m-l)\,\Phi(\mathcal{A}_2) \geq l\,\Phi_K(\varphi_L(X)) + (m-l)\,\Phi_K(\varphi_H(X)).$$

2. Пусть разбиение $\mathcal{A}_1$ оптимально на $\mathcal{A}_K(\varphi_L(X))$ и $\mathcal{A}(X) = \varphi_L^{-1}(\mathcal{A}_1)$, $\mathcal{A}_2 = \varphi_H(A)$. Тогда

$$m\,\Phi_K(X) \leq m\,\Phi(\mathcal{A}) = l\,\Phi(\mathcal{A}_1) + (m-l)\,\Phi(\mathcal{A}_2) \leq l\cdot\Phi_K(\varphi_L(X)) + (m-l)\,\Phi(\varphi_H(X)).$$

Теорема доказана.

Следствие I. Пусть $L = \{p_1, \dots, p_m\} \subset \{1, \dots, m\}$; $H_i = \{i\}$, $\varphi_i = \varphi_{H_i}$ для $i \notin L$. Тогда для любого K верно неравенство

$$0 \leq m\,\Phi_K(X) - l\,\Phi_K(\varphi_L(X)) \leq \sum_{i \notin L} \Phi(\varphi_i(X)).$$

Следствие 2. Пусть разбиение $\mathcal{A}_1$ оптимально на множестве $\mathcal{A}_K(\varphi_L(X))$ и $\mathcal{A}(X) = \varphi_L^{-1}(\mathcal{A}_1)$. Тогда

$$0 \leq \Phi(\mathcal{A}) - \Phi_K(X) \leq \frac{m-l}{m}\left[\Phi(\varphi_H(X)) - \Phi_K(\varphi_H(X))\right].$$

Теорема и следствия из нее дают возможность оценить степень неоднородности разбиения, оптимального на $\mathcal{A}_K(X)$, и построить приближенное решение задачи для множества признаков $T_1, \dots, T_m$ по степени неоднородности разбиения, оптимального на $\mathcal{A}_K(\varphi_L(X))$, т.е. по решению этой же задачи для признаков $T_{p_1}, \dots, T_{p_l}$.

Рассмотрим случай, когда по разбиению, оптимальному на $\mathcal{A}_K(\varphi_L(X))$, можно точно определить разбиение, оптимальное на $\mathcal{A}_K(X)$.

Теорема 2. Пусть $\mathcal{A}(X)$ – разбиение, оптимальное на множестве $\mathcal{A}_K(X)$, с такими центрами классов $C^1, \dots, C^K$, что для каждого $i \in H$ и $s = 1, \dots, K$ верно равенство $C_i^s = C_i^1$. Тогда для L -редукции $\tilde{X} = \varphi_L(X)$ выполняются условия:

I) разбиение $\mathcal{A}_1 = \varphi_L(\mathcal{A})$ оптимально на множестве $\mathcal{A}_K(\tilde{X})$;

2) для любого разбиения $\mathcal{A}_1'$, оптимального на множестве $\mathcal{A}_K(\tilde{X})$, разбиение $\mathcal{A}'(X) = \varphi_L^{-1}(\mathcal{A}_1')$ оптимально на множестве $\mathcal{A}_K(X)$.

Доказательство. Пусть $\mathcal{A}_1'$ – произвольное разбиение множества $\tilde{X}$, $\mathcal{A}' = \varphi_L^{-1}(\mathcal{A}_1')$, $\mathcal{A}_2 = \varphi_H(\mathcal{A})$ и $\mathcal{A}_2' = \varphi_H(\mathcal{A}')$. Наборы $\varphi_H(C^1), \dots, \varphi_H(C^K)$, по лемме I, являются центрами классов разбиения $\mathcal{A}_2$, и, по условию теоремы, для всех $s = 1, \dots, K$ имеем $\varphi_H(C^s) = \varphi_H(C^1)$. Тогда из лемм 3 и 4 получаем, что $\Phi(\mathcal{A}_2) = \Phi(\varphi_H(X)) \geq \Phi(\mathcal{A}_2')$. С другой стороны, $\Phi(\mathcal{A}) \leq \Phi(\mathcal{A}_2)$, откуда

$$m\,\Phi(\mathcal{A}) = l\,\Phi(\mathcal{A}_1) + (m-l)\,\Phi(\mathcal{A}_2) \leq l\,\Phi(\mathcal{A}_1') + (m-l)\,\Phi(\mathcal{A}_2') = m\,\Phi(\mathcal{A}').$$

Следовательно, $\Phi(\mathcal{A}_1) \leq \Phi(\mathcal{A}_1')$, т.е. разбиение $\mathcal{A}_1$ оптимально на множестве $\mathcal{A}_K(\tilde{X})$.

Если разбиение $\mathcal{A}_1'$ оптимально на $\mathcal{A}_K(\tilde{X})$, т.е. $\Phi(\mathcal{A}_1) = \Phi(\mathcal{A}_1')$, то из неравенства $\Phi(\mathcal{A}_2) \geq \Phi(\mathcal{A}_2')$ получаем

$$m\,\Phi(\mathcal{A}) = l\,\Phi(\mathcal{A}_1) + (m-l)\,\Phi(\mathcal{A}_2) \geq l\,\Phi(\mathcal{A}_1) + (m-l)\,\Phi(\mathcal{A}_2') = m\,\Phi(\mathcal{A}').$$

Следовательно, $\mathcal{A}'$ оптимально на множестве $\mathcal{A}_K(X)$.

Теорема доказана.

Л е м м а 5. Если разбиение $\mathcal{A}(X) = \{X_1, \ldots, X_K\}$ порождено наборами $a^1, \ldots, a^K \in P^m$ и для всех $s = 1, \ldots, K$ и $i \in H$ выполняется равенство $a^s = a^t$, то разбиение $\mathcal{A}_1 = \varphi_L(\mathcal{A})$ порождается наборами $\varphi_L(a^1), \ldots, \varphi_L(a^K)$.

Д о к а з а т е л ь с т в о. Условие леммы эквивалентно следующему утверждению: для любых целых s, t $(1 \le s, t \le K)$ и любого объекта $x^i \in X_s$ выполняются соотношения $\rho(x^i, a^s) \le \rho(x^i, a^t)$, $\varphi_H(a^s) = \varphi_H(a^t)$. Надо показать, что для любых целых s, t $(1 \le s, t \le K)$ и любого объекта $\tilde{x}^i \in \varphi_L(X_s)$ выполнено неравенство $\rho(\tilde{x}^i, \varphi_L(a^s)) \le \rho(\tilde{x}^i, \varphi_L(a^t))$. Пусть для произвольного фиксированного s $(1 \le s \le K)$ объект $\tilde{x}^i \in \varphi_L(X_s)$ и $x^i = \varphi_L^{-1}(\tilde{x}^i)$. Тогда, по определению множества $\varphi_L(X_s)$, имеем $x^i \in X_s$. Поэтому для любого $t = 1, \ldots, K$ выполняется неравенство $\rho(x^i, a^s) \le \rho(x^i, a^t)$. Отсюда получаем

$$m\rho(x^i, a^t) = l\rho(\tilde{x}^i, \varphi_L(a^t)) + (m-l)\rho(\varphi_H(x^i), \varphi_H(a^t)) =$$

$$= l\rho(\tilde{x}^i, \varphi_L(a^t)) + (m-l)\rho(\varphi_H(x^i), \varphi_H(a^s)).$$

Поэтому $\rho(x^i, a^t) - \rho(x^i, a^s) = \frac{l}{m}[\rho(\tilde{x}^i, \varphi_L(a^t)) - \rho(\tilde{x}^i, \varphi_L(a^s))]$.

Следовательно, $\rho(\tilde{x}^i, \varphi_L(a^s)) \le \rho(\tilde{x}^i, \varphi_L(a^t))$.

Лемма доказана.

Т е о р е м а 3. Пусть $\mathcal{A}(X)$ – разбиение, оптимальное на $\alpha_K(X)$ и с такими центрами классов $C^1, \ldots, C^K$, что для каждого $i \in H$ и $s = 1, \ldots, K$ верно равенство $C_i^s = C_i^1$. Пусть $\tilde{X} = \varphi_L(X)$ и некоторое разбиение $\tilde{\mathcal{A}}(\tilde{X})$ с центрами классов $\tilde{a}^1, \ldots, \tilde{a}^K$ оптимально на $\alpha_K(\tilde{X})$. Тогда любое разбиение $\mathcal{A}'(X)$, порожденное такими наборами $a^1, \ldots, a^K \in P^m$, что для всех $s = 1, \ldots, K$ верно $\varphi_L(a^s) = \tilde{a}^s$ и для всех $i \in H$ выполняется равенство $a^s = a^1$, является оптимальным на множестве $\alpha_K(X)$.

Д о к а з а т е л ь с т в о. Разбиение $\mathcal{A}_1' = \varphi_L(\mathcal{A}')$, по лемме 5, порождено наборами $\varphi_L(a^1), \ldots, \varphi_L(a^K)$, являющимися центрами классов разбиения $\tilde{\mathcal{A}}(\tilde{X})$, оптимального на $\alpha_K(\tilde{X})$. Из результатов § 2 вытекает, что разбиение $\mathcal{A}_1'$ также оптимально на множестве $\alpha_K(\tilde{X})$. Согласно теореме 2, получаем, что разбиение $\mathcal{A}_1 = \varphi_L(\mathcal{A})$ оптимально на множестве $\alpha_K(\tilde{X})$, следовательно, $\Phi(\mathcal{A}_1) = \Phi(\mathcal{A}_1')$. С другой стороны, так как центрами классов разбиения $\varphi_H(\mathcal{A})$ являются наборы $\varphi_H(C^1), \ldots, \varphi_H(C^K)$, а по условию теоремы для всех $s = 1, \ldots, K$ выполняется равенство $\varphi_H(C^s) = \varphi_H(C^1)$, то, по леммам 3 и 4, имеем

$$\Phi(\varphi_H(\mathcal{A})) = \Phi(\varphi_H(X)) \ge \Phi(\varphi_H(\mathcal{A}')).$$

Отсюда получаем

$$m\Phi(\mathcal{A}') = l\Phi(\mathcal{A}_1') + (m-l)\Phi(\varphi_H(\mathcal{A}')) \le$$
$$\le l\Phi(\mathcal{A}_1) + (m-l)\Phi(\varphi_H(\mathcal{A})) = m\Phi(\mathcal{A}).$$

Так как разбиение A , по условию теоремы, оптимально на множестве $\alpha_K(X)$, то разбиение A' оптимально на $\alpha_K(X)$.

Теорема доказана.

Поступила в ред.-изд.отдел.
12 февраля 1979 г.

Литература

1. Дюран Б., Оделл П. Кластерный анализ.-М.: Статистика, 1977. - 128 с.

2. Ту Дж., Гонсалес Р. Принципы распознавания образов.- М.: Мир, 1978. - 411 с.