а) все объекты из А, различные по системе признаков U, различались бы и по U';

б) если U_α и U_β из U', то

$$\sigma(\alpha,\beta) < E,$$

иначе –

$$\sigma(\alpha,\beta) > \delta;$$

в) U' была бы при этом минимальна по числу свойств (признаков).

4. Пусть задана мера сходства $\Lambda_{U^2}(i,j)$ между объектами a_i и a_j из А по совокупности свойств $U^2 = U/U'$ и заданы константы E и δ. Требуется разбить множество А по системе признаков U^2 на минимальное число групп A_1, A_2,... таких, что для всех объектов из А имело бы место

$$\Lambda_{U^2}(i,j) < E,$$

если a_i и a_j принадлежат разным образам, и

$$\Lambda_{U^2}(i,j) > \delta,$$

если a_i и a_j принадлежат одному и тому же образу.

5. Таким образом, алгоритм решения сложной задачи распознавания распадается на:

а) построение оптимальной совокупности U';

б) разбиение множества А на непересекающиеся группы по U^2;

в) отнесение неизвестного объекта к группе A_K (к = 1, 2, ...);

г) отнесение неизвестного объекта к образу A' или A'' внутри этой группы.

6. В вырожденном случае, когда в группах A_K содержится по одному объекту, задача не имеет решения.

7. Сказанное справедливо для любого числа образов.

А.Н. Дмитриев, А.А. Бишаев, В.О. Красавчиков,
Е.А. Смертин, Т.И. Штатнова

РАСПОЗНАВАНИЕ НА БАЗЕ ПОСТРОЕНИЯ ВСЕХ ТУПИКОВЫХ ТЕСТОВ

Пусть $\{X_1, \ldots, X_n, X_{n+1}\}$ – система признаков. Набор признаков $E = \{X_1, \ldots, X_n\}$ организует признаковое пространство, причем признаки $X_i, i = 1, \ldots, n$ называются характеристическими. Признак X_{n+1} фиксируется в качестве целевого.

Предполагается также, что признаковое пространство составлено из логических признаков, т.е. принимающих значения "истина" и "ложь". Целевой признак может быть как логическим, так и количественным (например, запасы нефти и газа). Имеет-

ся совокупность классов объектов $K_1, \ldots, K_M$, где M – число классов. Каждый объект класса K_j, $j = 1, \ldots, M$ охарактеризован всеми признаками, составляющими признаковое пространство. Обозначим символом "1" наличие и символом "0" – отсутствие у объекта того или иного признака из E.

В этих обозначениях каждый объект рассматриваемых классов представляется строкой нулей и единиц, а каждый класс K_j – таблицей T_j, $j = 1, \ldots, M$. Такие таблицы называются бинарными.

Определение 1. Бинарная таблица T называется допустимой, если все ее строки попарно различны.

Определение 2. Набор столбцов $t = t_{i1}, \ldots, t_{ik}$ допустимой таблицы T называется тестом, если при удалении из T всех столбцов, не входящих в t, полученная таблица является допустимой.

Определение 3. Тест называется тупиковым, если из него нельзя удалить ни одного столбца без того, чтобы он перестал быть тестом.

Определение 4. Пусть $A = \{T_1, \ldots, T_j, \ldots, T_M\}$ – набор бинарных таблиц с n столбцами в каждой. Если никакая строка $S = (\alpha_1, \ldots, \alpha_n)$ не входит одновременно в две таблицы из A, то набор A называется допустимым.

Определение 5. Набор признаков $t^* = \{x_{i1}, \ldots, x_{ik}\}$ назовем тестором допустимого набора таблиц $T_1, \ldots, T_M$, если при удалении из них всех столбцов, отличных от $i_1, \ldots, ik$, получим снова допустимый набор таблиц.

Определение 6. Тестор t^* называется тупиковым, если из него нельзя удалить ни одного признака без того, чтобы он перестал быть тестором.

Задача № 1. $M = 1$. Задан класс изучаемых объектов, представленный бинарной допустимой таблицей T. Рассматривается ситуация с неизвестными значениями целевого признака x_{n+1}. Предполагается что признаковое пространство сформировано в соответствии с целеуказанием, т.е. выбраны признаки, содержащие некоторую информацию о признаке x_{n+1}. Практически эти признаки являются косвенными, а их малая информативность компенсируется их количеством (детальностью описания).

Требуется: 1) упорядочить изучаемые объекты, представленные строками таблицы T, в соответствии с проявленностью целевого признака x_{n+1}; 2) отобрать признаки, наиболее существенные для данного класса объектов, и упорядочить признаковое пространство по существенности признаков.

Поставленная задача решается на базе построения всех тупиков тестов таблицы. Определим следующие величины:

1) $K(T)$ – число тупиковых тестов таблицы T;

2) $K_i(T)$ – число тупиковых тестов, в которые входит i-тый столбец.

Пусть j-тая строка таблицы T имеет вид $S_j = (t_{1j}, t_{2j}, \ldots, t_{nj})$. Основополагающими величинами для решения поставленной задачи являются:

а) $P(i) = \frac{K_i}{K}$ – "информационный вес i-того признака";

б) $J_S(j) = \sum_{i=1}^{n} t_{ij} P(i)$ – "информационный вес j-той строки".

Информационные веса признаков используются для классификации признаков по их существенности в отношении целей исследования таблиц. Информационные веса строк служат для упорядочения объектов по степени проявленности целевого признака.

Задача № 2. $M = 2$. Заданы 2 класса изучаемых объектов, представленные допустимым набором таблиц $\{T_1, T_2\}$, имеется объект, охарактеризованный строкой нулей и единиц $S = (\alpha_1 \ldots, \alpha_n)$. Требуется данный объект отнести к одному из двух классов, представленных таблицами T_1, T_2.

Задача решается на базе построения всех тупиковых тесторов допустимого набора таблиц. Определим следующие величины для набора $A = \{T_1, T_2\}$:

1) $K_i(A)$ – количество всех тупиковых тесторов системы таблиц A;

2) $K_i(A)$ – количество тупиковых тесторов, в которые входит i-тый признак;

3) $R_i = \frac{K_i(A)}{K(A)}$ – "разделяющий вес" i-того признака.

Пусть таблица T_1 имеет K строк, таблица T_2 – ℓ строк. Рассмотрим произвольную строку S длины n, составленную из единиц и нулей. Сопоставим этой строке и таблицам T_1 и T_2 следующие величины:

1) $R_1, R_2, \ldots, R_n$ – разделяющие веса признаков;

2) $P_1(1), P_1(2), \ldots, P_1(n)$ – информационные веса признаков по таблице T_1;

3) $P_2(1), P_2(2), \ldots, P_2(n)$ – информационные веса признаков по таблице T_2.

Пусть α_{ij} $(i = 1, \ldots, n;\ j = 1, \ldots, k)$, β_{ij} $(i = 1, \ldots, n;\ j = 1, \ldots, \ell)$ – элементы таблиц T_1 и T_2 соответственно и $S = (\gamma_1, \ldots, \gamma_n)$.

Определим операции $A \sim B$, $A \circ B$

$A \sim B$

B \ A	0	1
0	1	0
1	0	1

$A \circ B$

B \ A	0	1
0	0	1
1	1	0

Положим

$$1^{\circ}.\ \rho_1 = \frac{\sum_{i=1}^{n}\sum_{j=1}^{R} R(i)(\alpha_{ij} \oplus \gamma_i)}{R},$$

$$\rho_2 = \frac{\sum_{i=1}^{n}\sum_{j=1}^{\ell} R(i)(\beta_{ij} \circ \gamma_i)}{\ell};$$

$$2^{\circ}.\ \bar{\rho}_1 = \frac{\sum_{i=1}^{n}\sum_{j=1}^{R} P_1(i)(\alpha_{ij} \sim \gamma_i)}{R},$$

$$\bar{\rho}_2 = \frac{\sum_{i=1}^{n}\sum_{j=1}^{\ell} P_2(i)(\beta_{ij} \sim \gamma_i)}{\ell};$$

$$3^{\circ}.\ (\rho_{min})_1 = \min_j \sum_{i=1}^{n} R(i)(\alpha_{ij} \oplus \gamma_i),$$

$$(\rho_{min})_2 = \min_j \sum_{i=1}^{n} R(i)(\beta_{ij} \circ \gamma_i);$$

$$4^{\circ}.\ (\bar{\rho}_{min})_1 = \min_j \sum_{i=1}^{n} P_1(i)(\alpha_{ij} \circ \gamma_i),$$

$$(\bar{\rho}_{min})_2 = \min_j \sum_{i=1}^{n} P_2(i)(\beta_{ij} \circ \gamma_i).$$

Соответственно пунктам 1°–4° формируются 4 алгоритма распознавания.

1. Для таблицы T вычисляются величины $R_{(i)}$, экзаменуемая строка относится к первому классу при $\rho_1 < \rho_2$ и ко второму, если $\rho_1 > \rho_2$. При $\rho_1 = \rho_2$ распознавание строки S не проводится (отказ).

2. Вычисляются $\bar{\rho}_1$ и $\bar{\rho}_2$. Если $\bar{\rho}_1 > \bar{\rho}_2$, эталон S относится к первому классу, если $\bar{\rho}_1 < \bar{\rho}_2$ – ко второму. При $\bar{\rho}_1 = \bar{\rho}_2$ опознавание не производится.

3. Вычисляются $(\rho_{min})_1$ и $(\rho_{min})_2$. Если $(\rho_{min})_1 > (\rho_{min})_2$, то S относится ко второму классу, если $(\rho_{min})_1 < (\rho_{min})_2$ – к первому. При $(\rho_{min})_1 = (\rho_{min})_2$ опознавание S не происходит.

4. Строится аналогично 3, но с величинами $(\bar{\rho}_{min})_1, (\bar{\rho}_{min})_2$

При решении конкретных задач применяются и процедуры, использующие другие тестовые и тесторные параметры. Эти про-

цедуры возникают как в результате адаптации описанных выше алгоритмов, так и в связи с дальнейшей теоретической разработкой рассматриваемого круга вопросов.

В.И. Демин, Е.В. Кронгардт

ПОСЛЕДОВАТЕЛЬНЫЕ МЕТОДЫ РАСПОЗНАВАНИЯ ОБРАЗОВ

Выбор решающего правила является важнейшей составной задачей проблемы распознавания образов. Аппарат распознавания образов используется в настоящее время в ряде задач, связанных с прогнозом продуктивности локальных поднятий.

Вероятностно-статистические методы отвечают существу вопроса и имеют серьезную математическую базу. Минимальные предположения в этом случае заключаются в том, что множество объектов каждого образа является выборкой из генеральной совокупности с некоторой функцией распределения, имеющей плотность.

Рассмотрим случай двух образов.

Пусть $F(x)$ и $G(x)$ – гипотетические функции распределения, соответствующие первому и второму образу; $f(x)$ и $g(x)$ – плотности распределения; x – вектор произвольной размерности.

В случае равных априорных вероятностей классов и цен ошибок критические области, дающие минимум среднего числа ошибок, определяются неравенствами $\frac{f(x)}{g(x)} \leq 1$ и $\frac{f(x)}{g(x)} > 1$. По величине отношения $\frac{f(x)}{g(x)}$ делаем вывод о принадлежности объекта x к одному из классов (образов). По определению плотностей

$$\frac{f(x)}{g(x)} = \frac{\frac{\partial F(x)}{\partial x}}{\frac{\partial G(x)}{\partial x}}.$$

Заменяя производные дифференциалами с одинаковыми приращениями аргументов, получаем приближенное равенство

$$\frac{f(x)}{g(x)} \approx \frac{dF(x)}{dG(x)},$$

которое продолжим, заменив теоретические функции распределения эмпирическими,

$$\frac{f(x)}{g(x)} \approx \frac{dF_n(x)}{dG_n(x)}.$$

Величина, стоящая в правой части последнего равенства, есть отношение суммарного скачка эмпирической функции распределения первого класса к величине суммарного скачка эмпирической функции распределения второго класса в некоторой окрест-