DISCRETE MATHEMATICS
BANACH CENTER PUBLICATIONS, VOLUME 7
PWN—POLISH SCIENTIFIC PUBLISHERS
WARSAW 1982

ОБ ОДНОМ МЕТОДЕ ПОСТРОЕНИЯ ОПТИМАЛЬНЫХ КЛАССИФИКАЦИЙ

А. И. ЗЕНКИН, А. А. ЗЕНКИН

Вычислительный центр АН СССР, Москва, СССР

Рассматривается задача классификации, являющаяся одной из центральных в теории распознавания и прогнозирования. В работе предложен метод разбиений произвольных множеств объектов на классы в предположении, что априорное описание этих классов и их обучающие выборки — отсутствуют. Полученная классификация представляет собой решение некоторой задачи минимизации определенного типа функционалов, характеризующих качество классификации. Метод реализован в виде программного комплекса на языке АЛГОЛ-60 для ЭВМ БЭСМ-6.

§ 1. Вычисление расстояний между объектами

Пусть R — множество всех действительных чисел и T — прямоугольная матрица (таблица чисел из R) фиксированной размерности $M \times N$:

$$T = \{t_{ij}\}, \quad t_{ij} \in R, \quad i = 1, 2, \ldots, M, \quad j = 1, 2, \ldots, N. \tag{1}$$

С таблицей T однозначно связаны упорядоченная последовательность её строк

$$S = \{s_1, s_2, \ldots, s_M\} \tag{2}$$

такая, что

$$\forall i\big((s_i \in S) \to (s_i = (t_{i1}, t_{i2}, \ldots, t_{iN}))\big), \tag{3}$$

и последовательность её столбцов

$$S^+ = \{s_1^+, s_2^+, \ldots, s_N^+\} \tag{2a}$$

такая, что

$$\forall j\big((s_j^+ \in S^+) \to (s_j^+ = (t_{1j}, t_{2j}, \ldots, t_{Mj}))\big). \tag{3a}$$

Строки (2) таблицы T будем также называть *объектами* (или *векторами*), а элементы строк (3) — *признаками объекта* (или *компонентами вектора*).

Таким образом, задание числовой таблицы T (1) равносильно заданию упорядоченного множества S из M объектов (2), каждый из которых характеризуется набором из N признаков (3).

Для характеристики ,,близости'', ,,сходства'' или расстояния между строками исходной таблицы T введём ряд оценок.

1. *Оценки расстояний между строками.* Введём в рассмотрение N-мерный вектор весов признаков

$$\bar{w} = \bar{w}(w_1, w_2, \ldots, w_N), \quad 0 \leqslant w_j \leqslant 1,\ j = 1, 2, \ldots, N, \tag{4}$$

так что w_j — вес j-го признака таблицы T.

Пусть s_p, $s_q \in S$ — две произвольные строки из (2). Исходя из ,,физического'' смысла рассматриваемого ниже класса задач, за меру близости этих двух строк по j-ому признаку удобно взять некоторую функцию от абсолютной величины разности значений этого признака для объектов s_p и s_q:

$$r_j(p, q) = w_j f(|t_{pj} - t_{qj}|). \tag{5}$$

Следуя методу вычисления оценок, описанному в [1], введём числовой N-мерный вектор

$$\bar{\varepsilon} = \bar{\varepsilon}(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_N), \quad 0 \leqslant \varepsilon_j \leqslant 1,\ j = 1, 2, \ldots, N, \tag{6}$$

— вектор пороговых оценок по признакам, и определим явно функцию (5) с помощью следующего условия

$$r_j(p, q) = \begin{cases} 0, & \text{если} \quad |t_{pj} - t_{qj}| > \varepsilon_j, \\ w_j, & \text{если} \quad |t_{pj} - t_{qj}| \leqslant \varepsilon_j. \end{cases} \tag{7}$$

Расстояние $R(p, q)$ между строками s_p и s_q определим теперь следующим образом

$$R(p, q) = \frac{1}{N} \sum_{j=1}^{N} r_j(p, q). \tag{8}$$

Из (8) следует, что для любых s_p, $s_q \in S$ имеет место неравенство

$$0 \leqslant R(p, q) \leqslant 1. \tag{8a}$$

Введём пороговую оценку ,,близости'' строк

$$0 \leqslant \delta_S \leqslant 1 \tag{8б}$$

и на множестве всех пар объектов из (2) определим бинарную функцию ,,близости'' для строк по порогу δ_S:

$$\Gamma(p, q) = \begin{cases} 0, & \text{если} \quad R(p, q) < \delta_S, \\ 1, & \text{если} \quad R(p, q) \geqslant \delta_S. \end{cases} \tag{9}$$

Бинарная функция (9) имеет следующую содержательную интерпретацию: если усреднённое количество ,,совпадающих'' по ε-порогу признаков больше порога δ_S, то строка s_p ,,голосует'' за ,,сходство'' со строкой s_q (или, что то же, строка s_q ,,голосует'' за ,,сходство'' со строкой s_p); в противном случае количество ,,голосов'' за ,,сходство'' строк s_p и s_q по порогу δ_S равно нулю.

2. *Оценка расстояния от строки s_p до множества K.* Пусть $K \subset S$ — некоторое подмножество строк из (2)

$$K = \{s_{1'}, s_{2'}, \ldots, s_{M'}\}, \quad M' \leqslant M.$$

За расстояние от произвольного объекта $s_p \in S$ до заданного множества объектов K примем следующую величину:

$$R(p; K) = \frac{1}{M'} \sum_{s_q \in K} \Gamma(p, q), \tag{10}$$

которую можно интерпретировать как усредненное по числу объектов в K количество ,,голосов'' в пользу принадлежности рассматриваемого объекта s_p к множеству K.

Введём пороговую оценку

$$0 \leqslant \delta_K \leqslant 1 \tag{10a}$$

и на множестве всех подмножеств объектов из (2) определим бинарную функцию ,,близости'' для произвольной строки $s_p \in S$ и произвольного подмножества $K \subset S$ по порогу δ_K:

$$\Gamma(p; K) = \begin{cases} 0, & \text{если} \quad R(p; K) < \delta_K, \\ 1, & \text{если} \quad R(p; K) \geqslant \delta_K. \end{cases} \tag{11}$$

Введём теперь обобщенный вектор пороговых оценок

$$\bar{u} \underset{\text{Df}}{=} \bar{u}(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_N, \delta_S, \delta_K) = \bar{u}(\bar{\varepsilon}, \delta_S, \delta_K)$$

и следующее определение принадлежности объекта s_p множеству K объектов из S.

Определение 1. Объект $s_p \in S$ *принадлежит множеству объектов* K по фиксированному порогу $\bar{u}$, если $\Gamma(p; K) = 1$; в противном случае $s_p \notin K$.

§ 2. Статическая классификация
Задача классификации без учителя

Фиксируем вектор $\bar{u}(\bar{\varepsilon}, \delta_s, \delta_K)$ пороговых оценок и рассмотрим задачу о разбиении множества S последовательности строк (2) таблицы T на классы (число которых не задаётся априори).

Определение 2. *Классификацией* K упорядоченного множества объектов S по заданному вектору $\bar{u}(\bar{\varepsilon}, \delta_s, \delta_K)$ пороговых оценок назовём упорядоченную (вообще говоря, произвольно) последовательность

$$K = \{K_1, K_2, \ldots, K_L\} \tag{12}$$

подмножеств множества S, удовлетворяющую следующим условиям:

(1) $\forall i, j\big((i \neq j) \to (K_i \cap K_j = \varnothing)\big)$.

(2) $\bigcup_{i=1}^{L} K_i = S$.

(3) С точностью до порядка в (12) K не зависит от порядка строк в исходной таблице T.

(4) $\forall s_p \in S\Big((s_p \in K_i) \to \neg\, \exists j \neq i\big(R(p; K_j) > R(p; K_i)\big)\Big)$.

Условия (1) и (2) вытекают из определения понятия разбиения множества на классы; условие (3) исключает зависимость классификации от порядка (как правило, случайного) ввода исходной информации и означает следующее: если S' — любая последовательность, отличающаяся от S только порядком элементов, и $K' = \{K_{1'}, K_{2'}, \ldots, K_{L'}\}$ — классификация S' по тому же вектору $\bar{u}(\bar{\varepsilon}, \delta_s, \delta_K)$, то $L' = L$ и существует такое 1-1 соответствие φ между индексами последовательностей K и K', что

$$\forall i\big((i = \varphi(i')) \to (K_i \equiv K_{i'})\big), \quad i = 1, 2, \ldots, L.$$

Наконец, условие (4) можно содержательно интерпретировать как устойчивость классификации K по принадлежности любой строки ,,своему'' классу или, другими словами, для классификации K понятие ,,объект $s_p \in K_j$'' и ,,класс K_j является ближайшим (в смысле (11)) к объекту s_p'' должны совпадать.

Классификацию, удовлетворяющую введенному определению, мы будем строить в два этапа. Любую строку $s_p \in S$ будем называть в процессе классификации ,,отнесенной'', если она уже отнесена к какому-либо классу, и ,,неотнесенной'' — в противном случае. Очевидно, что до начала классификации любая строка $s_p \in S$ является неотнесенной, а после окончания классификации — отнесенной.

На I этапе строим так называемую предварительную классификацию

$$\tilde{K} = \{\tilde{K}_1, \tilde{K}_2, \ldots, \tilde{K}_{\tilde{L}}\}, \tag{13}$$

согласно следующему простейшему алгоритму.

Заносим в I класс строку s_1 и все строки $s_q \in S$, удовлетворяющие условию (9):

$$\Gamma(1; q) = 1. \tag{14}$$

Очевидно, что условие (14) однозначно определяет некоторый класс $\tilde{K}_1$. Если после этого в (12) имеются неотнесенные строки, то неотнесенную строку с наименьшим индексом, — пусть это будет, например, s_{i_2}, — и все неотнесенные строки $s_q \in S$, удовлетворяющие условию:

$$\Gamma(i_2, q) = 1,$$

звносим во II класс и формируем $\tilde{K}_2$. Эта процедура повторяется до тех пор, пока в (2) не останется неотнесенных строк. В результате мы получим предварительную классификацию $\tilde{K}$.

Очевидно, что классификация $\tilde{K}$ существенно зависит от порядка элементов в (2) и, следовательно, не удовлетворяет условию (3) определения 2. Для исключения зависимости классификации от порядка строк в исходной таблице T проводится II этап классификации представляющий собой следующую итерационную процедуру ,,перераспределения'' элементов последовательности (2) между классами (13).

Для каждого элемента $s_p \in S$ определяется

$$\max\{R(s_p; \tilde{K}_1), R(s_p; \tilde{K}_2), \ldots, R(s_p; \tilde{K}_{\tilde{L}})\}. \tag{15}$$

Пусть этот максимум достигается на $\tilde{K}_j$. Если $R(s_p; K_j) \geqslant \delta_K$, то строку s_p относим к классу $\tilde{K}_j$. При этом, если строка s_p принадлежала классу $\tilde{K}_{j'}$, и $j' \neq j$, то последний может оказаться пустым и будет в таком случае элиминирован. Эта процедура повторяется для всех элементов последовательности (2) до тех пор, пока не будет получена (точная) классификация

$$K = \{K_1, K_2, \ldots, K_L\}, \quad L \leqslant L', \tag{16}$$

обладающая тем свойством, что

$$\forall s_p \in S\big((s_p \in K_j) \to \max_{K_i \in K}\{R(s_p; K_i)\} = R(s_p; K_j)\big). \tag{17}$$

Поскольку в данной работе излагается лишь прикладной аспект разработанного метода классификации, заметим, что описанная процедура сходится очень быстро (для весьма различной по своему характеру исходной информации число итераций, как правило, не превышало 5).

§ 3. Динамическая классификация
Задача поиска оптимальной классификации

До сих пор рассматривалась классификация при фиксированном векторе $\bar{u}(\bar{\varepsilon}, \delta_S, \delta_K)$ пороговых оценок и было показано, что в этих условиях удаётся получить классификацию K, оптимальную в том смысле, что она не зависит от способа упорядочения классифицируемых объектов и является устойчивой с точки зрения принадлежности объектов определенным классом. Очевидно, что в общем случае

$$K = K(\bar{u})$$

и для различных $\bar{u}$ мы будем получать различные классификации. Поэтому возникает задача поиска классификаций, оптимальных в более широком смысле.

Назовём *допустимым* разбиение K множества объектов S на подмножества

$$K = \{K_1, K_2, \ldots, K_L\}, \tag{18}$$

если (18) удовлетворяет условиям (1) и (2), определения 2. Множество всех допустимых разбиений обозначим через $\mathscr{K}$. Пусть, далее, U — пространство допустимых значений вектора $\bar{u}$ пороговых оценок, определяемое неравенствами (6), (8б) и (10а).

Определим на $\mathscr{K}$ некоторый функционал $\mathscr{I}$, сопоставляющий любой допустимой классификации $K \in \mathscr{K}$ некоторое число $\mathscr{I}(K)$ и рассмотрим следующую задачу.

Для заданных начального ,,состояния'' $K(0) \in \mathscr{K}$ и начального управления $\bar{u}(0) \in U$ найти такое допустимое управление $\bar{u}(t)$, $t = 0, 1, \ldots$, которое является решением задачи

$$\mathscr{I} = \min_{\bar{u} \in U} \mathscr{I}\left(K(\bar{u})\right). \tag{19}$$

Отметим две особенности (и вытекающие из них трудности) указанной задачи. Во-первых, $\mathscr{K}$ — дискретное неупорядоченное множество и поэтому понятие траектории в $\mathscr{K}$ не имеет ,,физического'' смысла и, в лучшем случае, сводится к апостериорному упорядочению некоторых элементов из $\mathscr{K}$ в виде последовательности

$$K\left(\bar{u}(0)\right), K\left(\bar{u}(1)\right), \ldots, K\left(\bar{u}(i)\right), \ldots$$

Во-вторых, очевидно, что функционал $\mathscr{I}$ является ступенчатой функцией от $\bar{u}(t)$, что исключает возможность использования классических процедур минимизации. В связи с этим нами был разработан метод минимизации ступенчатых функций, в основу которого был положен метод нелинейного программирования, впервые описанный в [5] (и независимо разработанный в [2]).

На основе программного комплекса АСУМ МС [2], [3] была разработана и реализована на ЭВМ БЗСМ-6 подсистема для решения задачи (19). В частности, задача (19) была решена для следующих типов функционалов.

1. Во многих задачах априори известно количество L^* классов. В этом случае в качестве функционала задачи (19) берется выражение

$$\mathscr{I}(\bar{u}) = |L^p\left(K(\bar{u})\right) - L^*|, \tag{20}$$

где $L^p\left(K(\bar{u})\right)$ — ,,рассчитанное'' количество классов для классификации $K(\bar{u})$. Минимизируя функционал (20), мы найдём классификацию K^{opt}, содержащую L^* классов. Очевидно, что при столь общем критерии трудно гарантировать единственность решения задачи (19), тем не менее, если варьировать значение начального управления $\bar{u}(0)$, можно построить ряд классификаций

$$K_1, K_2, \ldots,$$

содержащих одинаковое число L^* классов, и использовать полученную информацию для принятия решения о выборе направления дальнейших исследований.

2. Из ,,физических'' соображений очевидно, что в общем случае качество классификации тем лучше, чем меньше расстояния между объектами внутри классов и чем больше расстояния между самими классами. Учитывая связь

$$\sum_{K_i \in K} \sum_{\substack{s_p \in K_i \\ s_q \in K_i \\ p \neq q}} R(p,q) + \sum_{\substack{K_i \in K \\ K_j \in K \\ i \neq j}} \sum_{\substack{s_p \in K_i \\ s_q \in K_j}} R(p,q) = \sum_{\substack{s_p \in S \\ s_q \in S \\ p \neq q}} R(p,q) = \text{const}. \tag{21}$$

для поиска классификации, оптимальной в указанном смысле достаточно использовать следующий функционал

$$\mathscr{I}(\bar{u}) = \sum_{K_i \in K(\bar{u})} \sum_{\substack{s_p \in K_i \\ s_q \in K_i}} R(p,q). \tag{22}$$

Решая задачу (19) с функционалом (22), мы получаем классификацию K^{opt}, при которой сумма расстояний между объектами внутри классов является минимальной (при этом, в силу (21), суммарное расстояние между классами будет, очевидно, максимальным).

3. Очевидно, что наибольший вклад с точки зрения формирования некоторого класса дают те признаки объектов, которые являются в данном классе наиболее близкими по своим значениям.

Обозначим через w_{ij} — вес j-го признака в i-ом классе и положим, по определению,

$$w_{ij} = 1 - \sigma_{ij}/\Delta_j, \tag{23}$$

где σ_{ij} — средне-квадратическое отклонение значения j-го признака в i-ом классе, а Δ_j — величина абсолютного колебания значений j-го признака (значений столбца s_j^+) в таблице T.

Рассмотрим произвольную классификацию K. Пусть класс $K_i \in K$. Упорядочим все N признаков по убыванию весов (23) в классе K_i

$$w_{ij_1} \geqslant w_{ij_2} \geqslant \ldots \geqslant w_{ij_N}. \tag{24}$$

Введём в качестве порога целое число $\delta_D < N$ и следующее

Определение 3. Назовём *понятием* о классе K_i по порогу δ_D совокупность признаков с номерами

$$j_1, j_2, \ldots, j_{\delta_D}. \tag{25}$$

Другими словами, понятием о классе является совокупность δ_D признаков с наибольшими весами.

Мы построили классификацию K и, в частности, получили класс $K_i \in K$, используя все N признаков объектов. Если теперь для объектов класса K_i мы будем решать задачу распознавания (15) не по всем N признакам (что, в силу свойства (17) классификации K было бы тавтологией), а *по понятию* о классе K_i, т.е. только по совокупности признаков (25), то возможно, что некоторые объекты класса K_i будут отнесены к другим классам. Обозначим

количество таких „переотнесенных" объектов, точнее, их долю от общего числа объектов класса K_i — через ν_i и положим

$$\nu\big(K(\bar{u})\big) = \max_{K_i \in K(\bar{a})} \nu_i. \tag{26}$$

Решая задачу (19) с функционалом (26) мы получаем классификацию K^{opt}, оптимальную с точки зрения информативности совокупности понятий о классах этой классификации K^{opt}. Это позволяет, в общем случае, заменить в каждом классе описание объектов с помощью N признаков их описанием через понятие о классе, т.е. как правило, существенно меньшим числом признаков, что в конечном счете упрощает задачу распознавания новых объектов с помощью полученной классификации K^{opt}.

Отметим важность изучаемой задачи разбиения для проблемы прогнозирования, содержанием которой является установление связи между прогнозируемым параметром или, иначе говоря, состоянием объектов и переменными $\{x_i,\ i = 1, 2, \ldots, n\}$, определяющими эти состояния. При этом возникает необходимость в решении задачи оптимальной классификации множества реализаций контрольных выборок [4].

Литература

[1] Ю. И. Журавлев, В. В. Никифоров, Кибернетика 3 (1971), Киев.
[2] А. А. Зенкин, Канд. дисс., МГУ, Москва 1974.
[3] —, Доклады АН СССР 230 (1976), 1051.
[4] —, *О математических методах прогнозирования*, Знание, Москва 1976.
[5] [Р. Хуук, Т. А. Дживс] R. Hooke, T. A. Jeeves, J. Assoc. Comput. Mach. 8 (1961), 212.

Presented to the Semester
Discrete Mathematics
(February 15–June 16, 1977)

DISCRETE MATHEMATICS
BANACH CENTER PUBLICATIONS, VOLUME 7
PWN—POLISH SCIENTIFIC PUBLISHERS
WARSAW 1982

О ТОЧНОСТИ МЕТОДА „УСРЕДНЕНИЯ"

В. К. ЛЕОНТЬЕВ

Вычислительный центр АН СССР, Москва, СССР

Многие задачи теории кодирования, распознавания образов, экспертных оценок можно рассматривать как оптимизационные метрические задачи на единичном n-мерном кубе E^n. При этом стандартной метрикой обычно является метрика Эвклида или метрика Хэминга, а оптимизируемый функционал является линейным от некоторой функции этой метрики. Задача состоит в нахождении множества фиксированной мощности на котором этот функционал достигает своего максимума или минимума.

Хорошо известным методом нахождения верхних (или нижних) оценок функционалов такого рода является „метод случайного кодирования" или метод оценки экстремума функционала средним значением этого функционала, вычисленным по „подходящему" ансамблю подмножеств фиксированной мощности из E^n. Этот метод играет существенную роль не только в такого рода оптимизационных задачах, но и во многих других проблемах комбинаторного анализа.

Настоящая работа посвящена изучению точности метода „усреднения" для специального класса метрических оптимизационных задач, о котором шла речь выше. Основной результат работы состоит в следующем.

Введен класс функций R_0 для функционалов от которого „метод усреднения" даёт асимптотически точный результат. Показано, что класс R_0 является выпуклым конусом в пространстве всех вещественнозначных функций, определенных на $[0, +\infty]$. Описаны некоторые свойства этого конуса, а для подкласса выпуклых вниз и убывающих на $[0, +\infty]$ функций дано абстрактное описание подконуса R_0 в виде некоторого аналитического соотношения. Показано, что для функций из класса R_0 градиентный алгоритм, применённый к функционалам от этих функций приводит к асимптотически точному результату.

Пусть E^n — множество вершин единичного n-мерного куба или множество двоичных наборов длины n. Всюду в дальнейшем множество E^n мы будем рассматривать как n-мерное векторное пространство над полем из