АКАДЕМИЯ НАУК СССР

ИЗВЕСТИЯ

АКАДЕМИИ НАУК СССР

ТЕХНИЧЕСКАЯ КИБЕРНЕТИКА

(ОТДЕЛЬНЫЙ ОТТИСК)

6

МОСКВА · 1972

ИЗВЕСТИЯ АКАДЕМИИ НАУК СССР
ТЕХНИЧЕСКАЯ КИБЕРНЕТИКА
№ 6 1972

АЛГОРИТМ КЛАССИФИКАЦИИ МНОГОМЕРНЫХ ДИСКРЕТНЫХ ДАННЫХ

Ю. К. БЕЛЯЕВ, В. А. МАЛЫШЕВ, С. С. ФИЛИМОНОВА

(*Москва*)

1. Введение. К настоящему времени известно много разнообразных схем алгоритмов классификации, основанных на разных идеях. Ниже приводится алгоритм классификации, сочетающий приемы выделения характерных признаков [1, 2] со случайным поиском (как градиентным, так и дихотомическим) в пространстве решающих правил. Предполагается, что информация о классифицируемых объектах, записывается в виде n-мерного двоичного вектора. Сначала дается точное описание алгоритма, а затем говорится о его реализации, применении и возможных математических задачах, связанных с ним. Ставятся также задачи о вычислении вероятностных характеристик числа ложных признаков.

Пусть на множестве $\mathfrak{X}_n$ n-мерных двоичных векторов $x = (x_1, \ldots, x_n)$, $x_i = 0, 1$, $i = \overline{1, n}$, задано два, в общем случае неизвестных, распределения вероятностей P_A и P_B. $\mathfrak{X}_n$ с распределением P_A и P_B образует популяции A и B соответственно. Имеющаяся информация представлена двумя независимыми выборками $O_A{}^*$ и $O_B{}^*$ объемов N_A и N_B из популяции A и B. Результаты выборок можно записать в виде двух матриц $X_A = \|x^i_{Aj}\|$, $X_B = \|x^k_{Bj}\|$, $i = \overline{1, N_A}$, $k = \overline{1, N_B}$, $j = \overline{1, n}$, i-я (k-я) строка $x_A{}^i(x_B{}^k)$ матрицы $X_A(X_B)$ соответствует данным i-го (k-го) выбора из популяции $A(B)$, $x_c{}^r = (x_{c1}{}^r, \ldots, x_{cn}{}^r)$, $C = A, B$, $r = i, k$. По результатам выборок X_A, X_B требуется построить решающее правило, по которому новый вектор $x \in \mathfrak{X}$ будет отнесен к популяции A или B. Ограничимся рассмотрением только нерандомизированных решающих правил, для которых $\mathfrak{X}$ разбивается на две части $\mathfrak{X}_A \cup \mathfrak{X}_B = \mathfrak{X}$, $\mathfrak{X}_A \cap \mathfrak{X}_B = \phi$. $x \in \mathfrak{X}_A$ означает решение о включении x в популяцию A. B противном случае, когда $x \in \mathfrak{X}_B$, x относится к популяции B. Предлагаемый ниже алгоритм легко обобщить на случай, когда $\mathfrak{X}$ разбивается на три части $\mathfrak{X}_A$, $\mathfrak{X}_B$, $\mathfrak{X}_н$. При $x \in \mathfrak{X}_н$ решение о классификации не принимается.

Набор $\xi = (i_1, \ldots, i_k; \gamma_1, \ldots, \gamma_k)$ назовем признаком размерности k, $1 \leqslant i_1 < i_2 < \ldots < i_k \leqslant n$, $\gamma_i = 0, 1$. Признаку ξ сопоставляется множество $\mathfrak{X}_n(\xi) \subset \mathfrak{X}$ всех векторов, таких, что для $x \in \mathfrak{X}_n(\xi)$ $x_{ij} = \gamma_j$, $j = \overline{1, k}$. Введем характеристическую функцию множества $\mathfrak{X}_n(\xi)$: $f_\xi(x) = 1$, если $x \in \mathfrak{X}_n(\xi)$ (x обладает признаком ξ), $f_\xi(x) = 0$, $x \in \mathfrak{X}_n(\xi)$. Для заданных выборок $O_A{}^* \cup O_B{}^*$ каждому признаку ξ соответствуют два числа

$$k_A(\xi) = \sum_{x \in O_A} f_\xi(x), \qquad k_B(\xi) = \sum_{x \in O_B} f_\xi(x).$$

Эта пара чисел определяет точку $(k_A(\xi), k_B(\xi))$ в положительном квадранте на плоскости. Основная задача при построении алгоритма классификации состоит в выборе правила выделения в множестве признаков заданной размерности $k = 1, 2, \ldots$ двух непересекающихся подмножеств $\mathfrak{A}$ и $\mathfrak{B}$, называемых признаками классов A и B соответственно.

2. Алгоритм классификации. Если введенные выше подмножества признаков $\mathfrak{A}$ и $\mathfrak{B}$ сформированы, то решающее правило относит x к популяции A тогда и только тогда, когда

$$\sum_{\xi\in\mathfrak{A}} \varphi(\xi)\, f_\xi(x) \geqslant \sum_{\xi\in\mathfrak{B}} \varphi(\xi)\, f_\xi(x), \tag{2.1}$$

где $\varphi(\xi)$ — некоторая весовая функция, определенная на множестве признаков ξ. Ниже мы ограничимся случаем $\varphi(\xi) \equiv 1$, что соответствует кворумному решению по числу признаков, «голосующих» за A или B и лишь множествами $\mathfrak{A}$ и $\mathfrak{B}$, выделяемыми следующим способом. Пусть $\psi = (\psi_1^A, \psi_2^A, \psi_1^B, \psi_2^B)$, где $0 \leqslant \psi_i^A, \psi_i^B \leqslant 1$, $i = 1, 2$. Определим $\mathfrak{A} = \mathfrak{A}(\psi_1^A, \psi_2^A)$ как множество тех ξ, для которых

$$k_B(\xi) / k_A(\xi) \leqslant \psi_2^A, \quad k_A(\xi) / N_A \geqslant \psi_1^A.$$

Аналогично $\mathfrak{B} = \mathfrak{B}(\psi_1^B, \psi_2^B)$ — есть множество ξ, для которого

$$k_A(\xi) / k_B(\xi) \leqslant \psi_2^B, \quad k_B(\xi) / N_B \geqslant \psi_1^B.$$

Такой способ задания $\mathfrak{A}$ и $\mathfrak{B}$ представляется наиболее простым и естественным. Левые линейные неравенства предполагают появление достаточного числа признаков, «голосующих» за $A(B)$, по сравнению с числом признаков, «голосующих» за $B(A)$, а правые соответствуют появлению достаточно большого числа таких признаков.

Алгоритм состоит из нескольких описываемых ниже подпрограмм и связей между ними.

а) Выборка O_A^* случайным образом разбивается на две: обучающую O_A и контрольную R_A. Для этого, быть может, повторно используются случайные двоичные векторы $(\alpha_1, \ldots, \underset{N_A}{\alpha_{N_A}})$, $x_A{}^i \in O_A$, если $\alpha_i = 1$, и $x_A{}^i \in R_A$, если $\alpha_i = 0$ при условии, что $\sum_{i=1} \alpha_i = N_A{}^1$ удовлетворяют условию $|N_A{}^1 - (N_A - N_A{}^1)| < \sqrt{N_A}$. Аналогичным образом разбивается O_B^* на O_B и R_B. Дальнейшая работа алгоритма состоит из последовательных идентичных этапов.

б) На первом этапе для всех признаков $\xi = (i; \gamma)$ размерности 1 на основе частей обучающих выборок O_A и O_B подсчитываются числа k_A, k_B. В блок ячеек D_1 заносится таблица из $2n$ строк $\{(i, \gamma), k_A, k_B\}$.

в) Для заданной четверки чисел ψ, определяющей области $\mathfrak{A}$, $\mathfrak{B}$, в соответствии с (2.1) производится классификация векторов $x \in R_A \cup R_B$ и вычисляется η^k — процент правильно классифицированных векторов из контрольных выборок R_A, R_B (для первого этапа).

г) Поиск оптимального решающего правила производится путем последовательного выбора четверок чисел ψ, определяющих $\mathfrak{A}$ и $\mathfrak{B}$. Заметим, что рассматриваемый алгоритм поиска не накладывает существенных ограничений на выбор начального значения $\psi^{(1)}$. Значение четверки чисел ψ на k-м шаге итерационного процесса обозначается через $\psi^{(k)}$. На первом этапе положим $\psi^{(1)} = (x_1, x_1, 0, 0)$, $\psi^{(2)} = (1 - x_1, x_1, 0, 0)$, где $x_1 = \min\limits_{\xi\in D_1} (k_A / N_A, k_B / N_B)$. Сначала производится поиск оптимума по первой координате. Пусть уже построен вектор $\psi^{(i)} = (x_i, x_1, 0, 0)$. Тогда $\psi^{(i+1)}$ получается следующим образом. Если $\eta_i > \eta_{i-1}$, то при $x_i < x_{i-1}$ полагаем $x_{i+1} = x_i - \alpha_i$, если же $x_i \geqslant x_{i-1}$, то $x_{i+1} = x_i + \alpha_i$. В случае $\eta_i \leqslant \eta_{i-1}$ полагаем при $x_i < x_{i-1}$, что $x_{i+1} = x_{i-1} + \alpha_i$, а при $x_i \geqslant x_{i-1}$—, что $x_{i+1} = x_{i-1} - \alpha_i$. Здесь в соответствии с рекомендацией [3] $\alpha_i = \alpha_{i-1} / \tau$, $\alpha_1 = 1 / 1 + \tau$, $\tau = 1.62$. Если $\eta^{i+1} < \min\limits_i \max\limits_{i_1, i_2, i_3 \leqslant i} (\eta_{i_1}, \eta_{i_2}, \eta_{i_3}) + \delta$, то изменение первой координаты прекращается. Далее, аналогичным образом производится поиск по второй координате, а затем — по третьей и четвертой. Далее, все повторяется цикли-

чески. Алгоритм поиска производится по любой координате при фиксированных трех остальных. Например, при начале поиска по второй координате полагается $\psi^{(1)} = (x_1, x_2, x_3, x_4)$, $\psi^{(2)} = (x_1, 1 - x_2, x_3, x_4)$. Прекращение поиска определяется выбором числа δ, которое можно задавать с пульта (целесообразно сначала задавать $\delta = 100$, а затем, через заданное время счета, $\delta = 0\%$). После этого поиск прекращается. В качестве ответа на первом этапе выдается вектор $\psi^1 = (\psi_1{}^A, \psi_2{}^A, \psi_1{}^B, \psi_2{}^B)$ с максимальным значением η^1 и соответствующие ему признаки из таблицы D_1. Общее число признаков не должно превышать d_1.

Опишем индуктивно следующие этапы (k-й этап). После $(k-1)$-го этапа имеется таблица D_{k-1} отобранных признаков размерности $k-1$ и вектор $\psi^{k-1} = (\psi_1{}^A, \psi_2{}^A, \psi_1{}^B, \psi_2{}^B)$ с соответствующим ему η^{k-1}. Число признаков не превышает d_{k-1}. Выделяется группа ячеек D_k.

д) Для каждого признака $\xi_{k-1} = (i_1, \ldots, i_{k-1}; \gamma_1, \ldots, \gamma_{k-1})$ из таблицы D_{k-1} формируются все признаки вида $\xi_k = (i_1, \ldots, i_k; \gamma_1, \ldots, \gamma_k)$, где $i_k \neq i_1, \ldots, i_{k-1}$, $\gamma_k = 0, 1$. Аналогично а) последовательно перебираются полученные в д) признаки и записываются в таблицу D_k вместе с соответствующими им числами k_A, k_B. Занесение в D_k производится также индуктивно. Пусть уже часть признаков занесена в D_k. Новый признак ставится в D_k так, чтобы признаки были упорядочены по величине разностей $|k_A - k_B|$. При переполнении числа ячеек, отведенного для D_k, признаки с наименьшими значениями $|k_A - k_B|$ исключаются. Далее аналогично в) и г) производится поиск оптимального значения ψ^k.

Числа d_k следует задавать с пульта ЭВМ, так как, во-первых, это позволяет исключить переполнение оперативной памяти ЭВМ, а во-вторых, выбирать любое количество «самых лучших» признаков, и, наконец, менять стратегию отбора. Работу алгоритма можно останавливать с пульта.

3. Задачи и возможные применения. В связи с описанным выше алгоритмом возникает несколько сложных комбинаторных задач по подсчету характеристик совместного распределения чисел признаков с заданными областями значений k_A и k_B при различных предложениях на P_A и P_B.

Проверку статистической достоверности решающего правила классификации целесообразно проводить на двух независимых выборках, взятых из одной и той же популяции ($P_A = P_B$). В частности, можно предположить, что матрицы X_A, X_B, представляющие результаты двух выборок, образованы в результате реализации $n(N_A + N_B)$ взаимно независимых случайных величин

$$x_{A_j}{}^i, x_{B_j}{}^k = 0, 1, \qquad P\{X_{A_j}{}^i = 1\} = P\{X_{B_j}{}^k = 1\} = q_j, \quad j = \overline{1, n}.$$

Если размерность n соизмерима с объемами выборок N_A, N_B или больше их, то можно ожидать появления признаков, ложным образом дающих высокий процент классификации этих неразличимых данных. В частности, представляет интерес вычисление вероятностных характеристик (среднего числа, распределения и т. п.) числа ортодоксальных признаков. Признак ξ называем ортодоксальным, если одно из чисел $k_A = k_A(\xi)$ или $k_B = k_B(\xi)$ равно нулю. Полезно также получить распределение $\max\limits_{\xi} (k_A(\xi), k_B(\xi))$, где максимум берется по всем ортодоксальным признакам.

Решение указанных или аналогичных им задач дает определенную информацию о степени возможности появления ложных признаков в условиях обилия информации о небольшом числе объектов. Именно в такой ситуации находится исследователь при решении задачи поиска информативных признаков, дающих возможность индивидуального прогноза надежности изделий. Здесь в начальный момент времени замеряются впрок всевозможные физические параметры. Естественно ожидать, что большая часть их не будет связана с надежностью изделия. Аналогичная ситуация возникает и при прогнозе хода болезни в медицине.

Описанный выше алгоритм реализован в кодах машины «БЭСМ-3М» лаборатории статистических методов МГУ. Программа занимает I МОЗУ, используется магнитная лента, магнитный барабан, СП записи на ленту и барабан перевода и печати. Используются следующие ограничения N_A, $N_B \leqslant 235$, $d_k \leqslant 80$, $n \leqslant 90$. Программа предусматривает исключение любых координат с целью уменьшения размерности при поиске информативных параметров, а также исключение любых начальных данных во время счета.

В заключение заметим, что настоящий алгоритм применялся, в частности, для клинической диагностики наступления церебрального атеросклероза с психическими нарушениями и без них по результату двоичного кодирования историй болезни. Указанная работа проводилась на основе данных, собранных Чирковым А. М. и Шорниковым Б. С. Оказалось, что можно построить решающие правила прогноза появления и отсутствия психических нарушений с малым количеством признаков большой размерности или большим количеством признаков малой размерности с высоким процентом правильной классификации. В качестве другого примера применения отметим использование этого алгоритма для индивидуального прогноза надежности полупроводниковых элементов по начальным значениям физических параметров.

Поступило 25 XI 1971

ЛИТЕРАТУРА

1. Бонгард М. М. Проблема узнавания. «Наука», 1967.
2. Браиловский В. Л. Алгоритм распознавания объектов с многими параметрами и его приложения. Изв. АН СССР, Техническая кибернетика, 1964, № 2.
3. Уайльд Д. Дж. Методы поиска экстремума. «Наука», 1967.