МЕТОДЫ АНАЛИЗА ДАННЫХ
(Вычислительные системы)
1985 год Выпуск III

УДК 519.237

# АЛГОРИТМ МНОГОКЛАССОВОГО РАСПОЗНАВАНИЯ, ОСНОВАННЫЙ НА ЛОГИЧЕСКИХ РЕШАЮЩИХ ФУНКЦИЯХ

Г.С.Лбов,Н.Г.Старцева

В работе предлагается алгоритм LRP построения решающего правила распознавания по обучающей выборке, использующий класс логических решающих функций от разнотипных признаков [1]. Алгоритм строит решающее правило в виде дихотомического дерева и предназначен для распознавания контрольной выборки (распознавания объектов, не входящих в обучение). Допускаются пропуски некоторых значений признаков в обучающей и контрольной выборках.

## 1. Основные определения

Под деревом решений понимается корневое дихотомическое дерево, у которого каждой внутренней вершине (узлу) ставится в соответствие некоторый предикат, ветвям, исходящим из внутренней вершины, соответствует истинность или ложность высказывания, получающегося при замене признаков их значениями; конечным вершинам приписываются имена образов из множества $\Omega = \{1,\ldots,\omega,\ldots,k\}$, где $k$ - количество образов ($k \geq 2$).

При построении дерева решений осуществляется последовательное "наращивание" вершин дерева в соответствии с принципом присоединения "лучшей" вершины к "лучшей". Ясно, что такая направленная процедура перебора, вообще говоря, дает приближенное решение $R_M$ (через $R_M$ обозначено дерево решений с $M$ конечными вершинами) задачи построения оптимального дерева $R_M^0$. Под последним понимается дерево с $M$ конечными вершинами, на котором достигается минимальная вероятность ошибки, т.е.

$$P(R_M^0) = \min_{R_M \in \Phi_M} P(R_M) ,$$

где $\Phi_M$ – множество всевозможных решающих правил в виде дерева с конечными вершинами, которые используют признаки $X_1, \ldots, X_j, \ldots, X_n$.

Поскольку распределения вероятностей $\{P(\omega,x)\}$ неизвестны, дерево решений строится на основе обучающей выборки.

Для обучения задается некоторое множество объектов $A \subseteq \Gamma$, где $\Gamma$ – множество изучаемых объектов (генеральная совокупность). Этому множеству $A$ соответствует эмпирическая таблица $V = \{X_j(a_i)\}$, $a_i \in A$, $i = \overline{1,N}$, $j = \overline{1,n}$, где $N$ – число объектов множества $A$; $X_j(a)$ – значение признака $X_j$ для объекта $a_i$. Эмпирическую таблицу $V$ также будем называть обучающей выборкой. Через $D_j$ обозначим множество различных значений признака $X_j$, определенных на множестве $A$. Множество $D_j$ определяется в зависимости от типа $X_j$: $D_j$ – набор имен для номинального признака, $D_j$ – набор баллов для порядкового признака; для дискретно-количественного признака $D_j$ – набор значений; для количественного признака $D_j$ – интервал от $-\infty$ до $+\infty$.

Рассматриваются следующие типы предикатов: $J(a,E_j)$="$X_j(a) \in E_j$", где $a \in \Gamma$, $E_j \in W_j$. Для номинального признака под $W_j$ понимается множество различных значений признака $X_j$ или всевозможных объединений этих значений, содержащих не более трех элементов; для порядкового признака $W_j$ – множество различных баллов или объединение "соседних" баллов; для дискретно-числового признака $W_j$ – множество различных значений или объединение "соседних" значений; для количественного признака $W_j$ – множество интервалов типа $[\rho',\rho''] \cup [-\infty,\rho'']$, где $\rho',\rho'' \in D_j'$, а $D_j'$ – множество среднеарифметических значений двух соседних несовпадающих значений признака $X_j$; если две соседние точки $D_j'$ принадлежат одному образу, то среднеарифметическое значение между ними не рассматривается.

Для дискретно-количественных и количественных признаков предикат может иметь следующий вид: $J(a,E)$ = "$X(a) \in E$", где $X(a) = X_{j_1}(a), \ldots, X_{j_s}(a), \ldots, X_{j_m}(a)$, $E = \{x / \sum_{s=1}^{m} c_{j_s} \cdot X_{j_s}(a) > c_0\}$, $m = 2,3$, где $W$ – множество всевозможных подмножеств указанного типа, которые можно организовать на подсистеме признаков $X = \{X_{j_1}, \ldots, X_{j_s}, \ldots, X_{j_m}\}$. На множестве признаков $\{X_1, \ldots, X_j, \ldots, X_n\}$ можно сформулировать некоторое множество $\Theta$ предикатов указанных типов.

Конечной вершине $B^t$, $t = \overline{1,M}$, будет соответствовать конъюнкция такого типа: $S^t = S(a, \widehat{E}^t) = J(a, E^{1t}) \wedge J(a, E^{2t}) \wedge \ldots \wedge J(a, E^{\nu t}) \wedge \ldots \wedge J(a, E^{m_t t})$, где $J_{\nu t} = J(a, E^{\nu t})$ – один из указанных выше типов предикатов, $\widehat{E}^t = \prod_{\nu=1}^{m_t} E^{\nu t}$, $m_t$ – длина пути до вершины $B^t$, $M$ – количество конечных вершин, $l = \sum_{t=1}^{M} m_t$ – длина внешнего пути. Дерево решений задает разбиение $\alpha \in \Psi_M$ пространства признаков размерности $n$, т.е. $D = \bigcup_{t=1}^{M} E^t$, где $D = \prod_{j=1}^{n} D_j$, $E^t = \widetilde{E}^t \times \prod_{j \in I^t} D_j$, где $I^t$ – множество индексов тех признаков, которые не вошли в $S^t$, $\Psi_M$ – множество всевозможных разбиений в рассматриваемом классе.

## 2. Критерий качества дерева решений

На основе обучающей выборки для любого разбиения $\alpha \in \Psi_M$, заданного в виде дерева, для каждой конечной вершины $B^t$ может быть указано распределение количества объектов по образам, т.е. $(\mu_1^t, \ldots, \mu_\omega^t, \ldots, \mu_k^t)$, где $\mu_\omega^t$ – число объектов образа $\omega$; приписано решение $\omega^t$, где $\omega^t$ определяется из соотношения $\mu_{\omega_t}^t = \max_\omega \mu_\omega^t$; определено число ошибок $\widetilde{\mu}^t = \sum_{\substack{\omega=1 \\ \omega \neq \omega_t}}^{k} \mu_\omega^t$. В результате получаем выборочное решающее правило в виде дерева решений $R_M$.

Таким образом, дереву решений $R_M$ может быть сопоставлена оценка вероятности ошибки $P(R_M) = \sum_{t=1}^{M} \frac{\widetilde{\mu}^t}{N}$, вычисленная по обучающей выборке. Наилучшим деревом $R_M$ будем считать дерево, которое дает минимальную оценку вероятности ошибки, т.е. $\overline{P}(\overline{R}_M) = \min_{R_M \in \Phi_M'} \overline{P}(R_M)$, где $\Phi_M'$ – множество всевозможных деревьев, которые можно задать, используя указанные выше предикаты.

Из теоретических исследований [2] ясно, что, чем больше сложность класса решающих правил и меньше объем выборки, тем больше может быть отклонение оценки $\overline{P}(\overline{R}_M)$ от вероятности ошибки $P(\overline{R}_M)$.

Отсюда следует, что использование $\bar{P}(R_M)$ в качестве оценки критерия дерева не всегда является оправданным. Необходимо еще учитывать сложность класса решающих правил. В работе [I] показано, что при фиксированном объеме выборки $N$ и размерности пространства $n$ сложность класса логических решающих правил зависит от числа конечных вершин $M$ : чем больше $M$, тем больше может быть отклонение от оптимального правила (принцип минимального числа вершин). Отсюда следует, что при одинаковых оценках $\bar{P}(R_{M_1})$ и $\bar{P}(R_{M_2})$ необходимо выбирать дерево с минимальным числом $M$ . Сложность решающего правила зависит также от вида предиката.

В работе [3] дополнительно показано, что при прочих равных условиях среди различных деревьев решений с $M$ конечными вершинами оптимальным по Байесу будет то, в котором реализации из эмпирической таблицы $V$ распределены равномерно по конечным вершинам, т.е. вершине соответствует примерно равное число объектов $\mu^t \approx \frac{N}{n}$ (принцип равномерности).

В работе [4] также показано, что для уменьшения времени принятия решения необходимо выбирать дерево, минимизирующее длину внешнего пути $l$ (принцип минимизации пути).

Исходя из указанных замечаний, был сформулирован критерий оценки любой конечной вершины, отражающий перспективность ее ветвления в процессе построения дерева по обучающей выборке.

## 3. Описание алгоритма

Обозначим критерий оценки качества вершины через $F_r$, $r=\overline{1,R}$, где $R$ - число конечных вершин, полученных в текущий момент ($R = \overline{1,M}$, $M$ задано). Описание критерия приводится ниже. На первом шаге выбирается из множества возможных предикатов $\Theta$ лучший в смысле критерия $F_1$ предикат $J_1$ . Объекты обучающей выборки $V$ разбиваются на две группы: $V_{J_1}$ - объекты, для которых предикат $J_1$ истинен (объекты относятся к вершине $b_2$ ), и $V_{\bar{J}_1}$ - для которых $J_1$ ложен (объекты относятся к вершине $b_3$). На втором шаге для каждой группы объектов ищется свое, лучшее в смысле $F_2$ и $F_3$ высказывание $J_2$ и $J_3$ . Объекты разбиваются на четыре группы: $V_{J_1 J_2}$ - для которых $J_1$ и $J_2$ истинны; $V_{J_1 \bar{J}_2}$ - для которых $J_1$ истинен, $J_2$ ложен; $V_{\bar{J}_1 J_2}$ - для которых $J_1$ ложен, $J_2$ истинен; $V_{\bar{J}_1 \bar{J}_2}$ -

для которых $J_1$ и $J_2$ ложны. Деление объектов будет продолжено из вершины с наименьшим значением критерия из $\{F_2, F_3\}$ и т.д. На $r$-м шаге из множества внутренних вершин $\{b_1, \dots, b_r, \dots, b_R\}$ деление согласно "принципу минимального числа вершин" будет продолжено из вершины $b_r$, для которой значение критерия $F_r$ наименьшее, с учетом внешнего пути. Рассмотрим вершины $b_{r'}$ и $b_{r''}$. Пусть $F_{r'}$ – минимальное по разбиениям значение критерия для вершины $b_{r'}$, $F_{r''}$ – для вершины $b_{r''}$; если $|F_{r'} - F_{r''}| < \varepsilon$, то дальнейшее построение дерева будет происходить с учетом "принципа минимизации пути" из вершины, при ветвлении которой получается дерево с меньшей длиной внешнего пути $l$.

Ветвление продолжается до тех пор, пока не достигнем максимального количества вершин $q$, заданных в начале построения дерева (параметр $q$ – суммарное количество конечных и внутренних вершин), или если все вершины $b_r$ делению не подлежат, т.е. являются конечными.

Рассмотрим условие, при котором вершина не подлежит дальнейшему делению. Обозначим через $b_{r'}$ и $b_{r''}$ вершины, выходящие из узла $b_r$. Если существует $d_r$ $(1 \le d_r \le k)$ такое, что

$$\sum_{\substack{\omega=1 \\ \omega \ne d_r}}^{k} \mu_\omega^r < \alpha ,$$

где $\alpha$ – некоторый порог, то такая вершина $b_r$ является конечной, где $\mu_\omega^r$ – количество объектов образа $\omega$ в вершине $r$. Вершина $b_r$ подлежит дальнейшему делению, если существуют $d'$ и $d''$ $(1 \le d', d'' \le k)$ и предикат $J \subseteq \Theta$, такие, что $\mu_{d'}^{r'} > \beta$ и $\mu_{d''}^{r''} > \beta$, где $\beta$ – некоторый порог, $\mu_{d'}^{r'}$ – количество объектов образа $d'$ в вершине $b_{r'}$ при условии, что в вершине $b_r$ рассматриваемый предикат $J$ истинен, и $\mu_{d''}^{r''}$ – количество объектов в вершине $b_{r''}$ образа $d''$, если $J$ ложен. Каждой вершине $b_{r'}$ и $b_{r''}$ приписывается решение относительно образа $d_\nu$, согласно следующему правилу: $\mu_{d_\nu}^{\nu} = \max\limits_{\omega} \mu_\omega^r$, где $\nu = r', r''$.

Пусть $Q \subseteq \Theta$ – множество таких предикатов $J$, которые удовлетворяют условиям $\mu_{d'}^{r'} > \beta$ и $\mu_{d''}^{r''} > \beta$. Если $|Q| > 1$, то выбирается такой $J^* \in Q$, который минимизирует критерий $F_r$.

Критерий $F_r = F' + F''$, где

$$F' = \frac{\sum\limits_{\omega=1}^{k} \mu_\omega^r - \max\limits_\omega \mu_\omega^{r'} - \max\limits_\omega \mu_\omega^{r''}}{\sum\limits_{\omega=1}^{k} \mu_\omega^r}, \qquad F' \in [0,1],$$

$$F'' = \frac{\min\limits_\omega (\mu_\omega^{r'}, \mu_\omega^{r''})}{\sum\limits_{\omega=1}^{k} \mu_\omega^r}, \qquad F'' \in [0, \tfrac{k}{2}].$$

Величина $F'$ - относительное число ошибок распознавания объектов, соответствующих вершине $b_r$. Если $a_i$ - объект из вершины $b_r$, который нельзя отнести ни к одной из вершин $b_{r'}, b_{r''}$ (так как в объекте $a_i$ есть пропуск), то, как видно из определения $F'$, значение критерия увеличивается.

Добавка $F''$ позволяет учесть качество разделения объектов по всем $k$ образам, при $k = 2$ добавка не учитывается.

Необходимо отметить, что критерий $F_r$ тем меньше, чем больше суммарное число объектов по всем образам в вершине $b_r$ (реализует "принцип равномерности").

Если вершина $B^t$ $(t = \overline{1,M})$ конечна, то в ней задается решение относительно образа, как это было показано выше.

Окончательный вид критерия будет уточняться по мере проведения экспериментов.

Рассмотрим параметры $\alpha$, $\beta$, $\varepsilon$, $q$.

Параметр $\alpha$ - максимально-допустимое количество объектов, неверно распознанных в вершине $b_r$. Величина $\alpha$ зависит от количества заданных вершин $q$. Обычно фиксируется относительное максимально-допустимое число ошибок $\bar{P}_0$ на обучающей выборке, тогда

$$\alpha = \left[\frac{\bar{P}_0 N}{\frac{q}{2} + 1}\right].$$

Параметр $\beta$ - минимально-допустимое количество объектов в вершине $b_r$, принадлежащих образу $\omega$, по которому принимается решение. Величина $\beta \in [1, N^*]$, где $N^* = \min\limits_\omega N_\omega$ ($N_\omega$ - число объектов образа $\omega$).

Из сказанного ясно, что учет параметров $\alpha, \beta$ приводит к более равномерному распределению объектов обучающей выборки по конечным вершинам, тем самым реализуется "принцип равномерности".

Параметр $\varepsilon \in [0,2]$ обычно выбирается близким к 0.

Параметр $q \in [3, 2N-1]$. Как показывает опыт решения прикладных задач, достаточно хорошее решение получается при $q < 4k$.

Пусть решающее правило построено. Для распознавания объекта $a_i$ подставляются конкретные значения признаков в предикаты, находящиеся в узлах дерева, и достигается некоторая конечная вершина $B^t$, которой соответствует образ $\omega$. Объекту $a_i$ приписывается образ $\omega$.

Если значение предиката $J$ в вершине $b_r$ для объекта $a_i$ не определено ($x_j$ - признак, входящий в $J$, в объекте $a_i$ отсутствует), то проверяем два возможных пути из вершины $b_r$, где $J$ - истина и $J$ - ложь. Если встречаются $p$ неопределенных предикатов, то получается $(p+1)$-а конечных вершин $\{B_1, \ldots, B_{p+1}\}$ и объекту $a_i$ приписывается образ $d = \max\limits_{\omega} \left\{ \sum\limits_{s=1}^{p+1} \mu_1^s, \ldots, \sum\limits_{s=1}^{p+1} \mu_\omega^s, \ldots, \sum\limits_{s=1}^{p+1} \mu_k^s \right\}$.

Для распознанного объекта дополнительно вычисляется $\tilde{p}(\omega,t)$ - относительное число объектов образа $\omega$ в вершине $B^t$.

## 4. Описание линейного режима

Как было отмечено выше, в качестве предиката в узле дерева могут быть использованы высказывания типа $\sum\limits_{s=1}^{m} c_{i_s} x_{i_s}(a) > c_0$ ($m = 2,3$).

В этом случае в вершине $b_r$ в качестве условия деления рассматривается гиперплоскость для тех двух образов из $k$, для которых число объектов максимально в этой вершине на всем признаковом пространстве $D$ для количественных и дискретно-количественных признаков. Не рассматриваются признаки, в которых есть пропуски. Для простоты описания будем считать, что все $n$ признаков количественные и не имеют пропусков.

Дискриминантная функция имеет вид [5]:

$$\phi(X) = X' \bar{\Sigma}^{-1} (\bar{E}^{(1)} - \bar{E}^{(2)}) - $$
$$- \frac{1}{2} (\bar{E}^{(1)} + \bar{E}^{(2)})' \, \bar{\Sigma}^{-1} (\bar{E}^{(1)} - \bar{E}^{(2)}) - \ln \frac{p_2}{p_1}, \qquad (1)$$

где $p_\omega$ - оценки априорных вероятностей класса $\omega$ ($\omega = 1,2$), $E^{(\omega)} = (E_1^{(\omega)}, \ldots, E_\nu^{(\omega)}, \ldots, E_N^{(\omega)})$ - вектор средних значений образа $\omega$, определенных по выборке.

Если $n > 7$, то ковариационная матрица $\bar{\Sigma}$ представляется в виде диагональной и вычисляется по общей выборке рассматриваемых двух образов. В дискриминантной функции (I) остаются те семь признаков из исходной системы, у которых коэффициенты при $X_j$ имеют максимальные по модулю значения.

Если $n \leq 7$, то $\Sigma$ - ковариационная матрица, определенная по выборке рассматриваемых двух образов. Для всевозможных комбинаций по два и по три признака строится гиперплоскость (I). Каждая гиперплоскость разбивает объекты всех образов на две группы: лежащие слева от гиперплоскости $(\varphi(X) \leq 0)$, что эквивалентно истинности предиката, и лежащие справа - эквивалентно ложности предиката. Выбирается та гиперплоскость, для которой значение критерия $F_r$ минимально.

Алгоритм реализован на языке FORTRAN для машин типа ЕС. Описание программы приводится в [6].

## Литература


I. ЛБОВ Г.С. Методы обработки разнотипных экспериментальных данных.- Новосибирск: Наука, 1981. - 160 с.

2. ВАПНИК В.Н., ЧЕРВОНЕНКИС А.Я. Теория распознавания образов.- М.: Наука, 1974. - 415 с.

3. ДОНСКОЙ В.И. Алгоритмы обучения, основанные на построении решающих деревьев. -Журнал выч.математики и мат.физики,1982,т.22, № 4,с.963-974.

4. ДИСКАНТ В.А. Алгоритм построения правил классификации в структурно-аналитических моделях распознавания,- В кн.: Математические методы анализа динамических систем, Харьков,1983, вып.7, с.10-15.

5. АНДЕРСЕН Т. Введение в многомерный статистический анализ.- - М.: ФМ,1963. - 472 с.

6. СТАРЦЕВА Н.Г. LRP - логическое распознающее правило (Описание программы. Решение модельных и прикладных задач). - Настоящий сборник, с.11-22.



Поступила в ред.-изд.отд.
28 мая 1985 года