МЕТОДЫ АНАЛИЗА ДАННЫХ
(Вычислительные системы)
1985 год Выпуск III

УДК 519.237

LRP - ЛОГИЧЕСКОЕ РАСПОЗНАЮЩЕЕ ПРАВИЛО
(Описание программы. Решение модельных и прикладных задач)

Н.Г. Старцева

В работе даны краткое описание программы LRP, реализующей алгоритм LRP (см. [I]), иллюстрация работы программы на двух модельных примерах, а также сравнение LRP с четырьмя другими известными алгоритмами (на модельных и реальных примерах), которое позволяет убедиться в правильности принципов, заложенных в алгоритме LRP.

§ I. Программа LRP

Программа LRP (оформленная как подпрограмма) предназначена для построения решающего правила в виде логического дерева с заданным количеством вершин, соответствующих обучающей выборке (обучению), последняя (выборка) может быть разбита на K ($K \geq 2$) образов. С помощью этого правила распознается контрольная выборка (контроль).

Программа LRP позволяет работать с разнотипными данными и таблицами, в которых есть пропуски. При построении решающего правила LRP одновременно выделяет подсистему информативных признаков. Общие понятия и теоретические исследования так же, как и алгоритм LRP, описаны в [I]. В приложении I даются описание параметров работы LRP и обращение к программе.

Качество работы алгоритма можно определить с помощью приема "скользящий экзамен", реализованного в программе, который состоит в следующем: из обучающей выборки выделяют один элемент и предлагают обучиться на оставшейся части последовательности и классифицировать выделенный элемент, затем выделяют другой элемент (первый возвращается на место) и снова проводят обучение и экзамен на

этом элементе, и так поочередно перебирают все элементы обучающей последовательности. Затем подсчитывается, сколько раз алгоритм ошибался при классификации выделенных элементов. Отношением числа ошибочных классификаций к объему выборки и оценивается качество решающего правила [2].

Рассмотрим режимы работы LRP и параметры, регулирующие качество решающего правила.

В программе реализованы четыре основных режима работы, задаваемые с помощью параметра IR2: IR2=0 - обучение; IR2 = 1 - обучение и контроль; IR2 = 2 - скользящий экзамен; IR2 = 3 - обучение и скользящий экзамен.

В качестве условия деления объектов на каждой вершине дерева рассматриваются простые высказывания или строится гиперплоскость. В зависимости от вида условия деления различаются три режима, устанавливаемые параметром NRE: NRE = 1 - режим без гиперплоскости; NRE = 2 - режим с гиперплоскостью; NRE = 3 - режим только с гиперплоскостью.

В каждой вершине дерева ищется оптимальное в смысле некоторого критерия условие деления объектов. С помощью параметра IR можно выбирать тот или другой критерий деления. При IR = 1 в программе LRP реализован только один критерий [2]. При IR = 2,3,4,5 можно вставить подпрограмму-функцию, аналогичную реализованной в LRP , результатом которой является другой критерий деления (исследования по выбору новых критериев будут проведены позже).

Алгоритм устроен таким образом, что деление каждый раз происходит из той вершины, у которой значение критерия наименьшее (LD = = 0), но в LRP реализован также режим (LD = 1) , при которрм,если значения критериев для двух вершин отличаются меньше, чем на EPS (обычно EPS = 0), то деление будет происходить из вершины с наименьшей длиной внешнего пути.

Улучшать результаты обучения можно с помощью параметров KALF, KBET, KGAM, KWT, EPS. Параметры KALF = α, KBET = β, KWT = q, EPS = ε приведены в [I]. Через KGAM обозначено минимально возможное количество объектов каждого из двух классов,по которым строится гиперплоскость.

§ 2. Модельный пример и результаты решения

Количество образов в примере равно трем. Объем обучающей выборки - 20 реализаций (по 6 реализаций первого и второго образов

и 8 реализаций третьего образа), объем контрольной выборки - 4 реализации. В обучающей и контрольной выборках есть пропуски. Количество признаков - 6. Первый признак - наименование, четвертый - порядковый, остальные количественные.

Для второго и третьего признаков распределение строится следующим образом: первый образ представляет собой нормальное распределение $N(\mu_1, \sigma^2)$, второй - $N(\mu_2, \sigma^2)$, третий - смесь двух нормальных распределений $N(\mu_3, \sigma^2)$ и $N(\mu_4, \sigma^2)$ с априорными вероятностями $\frac{1}{4}$ и $\frac{3}{4}$ соответственно, где $\mu_1 = \begin{pmatrix} -1 \\ 1 \end{pmatrix}$, $\mu_2 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$, $\mu_3 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$, $\mu_4 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$, $\sigma = 0{,}62$.

Первый признак принимает имена "3" или "6" (с равной вероятностью), если при этом второй и третий признаки распределены по нормальному закону $N(\mu_4, \sigma^2)$, и имена "I" и "2" (с равной вероятностью) - для всех остальных реализаций. Четвертый признак - порядковый и равен "5" для реализаций, у которых второй и третий признаки распределены по нормальному закону $N(\mu_3, \sigma^2)$, и равен "3" или "4" (с равной вероятностью) для всех других реализаций. Пятый и шестой признаки неинформативные (шумовые) и распределены одинаково и независимо для трех образов - $N(0,20)$.

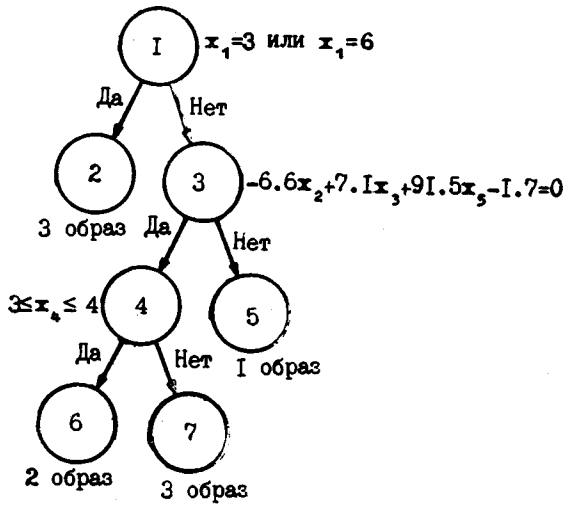

Рис.I

В приложении 2 приводится распечатка результатов работы программы.

На рис.I приведено дерево решений, построенное по результатам обучения.

Построенное решающее правило в виде дерева решений с нулевой ошибкой разбивает объекты обучающей выборки на три класса и является наглядной интерпретацией логических закономерностей данной моделируемой выборки.

§ 3. Результаты сравнения LRP с четырьмя алгоритмами распознавания образов на модельном примере

Количество классов в модельном примере равно двум. Объем обучающей выборки - 100 реализаций (по 50 реализаций каждого образа), объем контрольной выборки - 200 (по 100 реализаций каждого образа). Количество признаков - 20.

Первый признак распределен следующим образом: для первого класса признак попадает в интервалы (0,1), (2,3), (4,5) с вероятностью 0,95; в интервалы (1,2), (3,4), (5,6) с вероятностью 0,05 и распределен внутри каждого интервала равномерно. Для второго класса наоборот. Остальные 19 признаков неинформативные, они распределены одинаково и независимо для обоих классов $N(0,20)$.

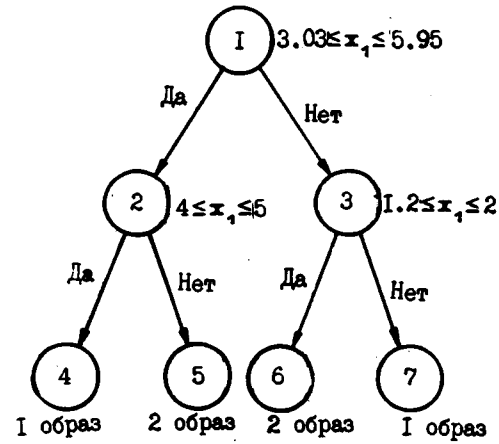

Рис.2

Рассмотрим пять алгоритмов распознавания образов: первый алгоритм использует дискриминантную функцию Фишера [3], второй - использует квадратичную дискриминантную функцию [3], третий основан на непараметрической оценке [4], четвертый - на логических функциях [5], пятый - LRP.

Качество алгоритма определяется оценкой вероятности ошибочной классификации на контроле - $\bar{P}_{\alpha}$, усредненной по числу экспериментов*), где α - номер алгоритма. Усреднение проводилось

*) Под экспериментом понимается: генерирование обучающей выборки, построение решающего правила, генерирование контрольной выборки и вычисление оценки вероятности ошибки для данного решающего правила и заданного контроля.

по девяти экспериментам. Результаты работы первых четырех алгоритмов на данном модельном примере уже приводились в [6]: $\bar{P}_1 = 0,45$, $\bar{P}_2 = 0,47$, $\bar{P}_3 = 0,47$, $\bar{P}_4 = 0,18$. Для алгоритма LRP оценка $\bar{P}_5 = 0,08$. Из приведенных выше результатов видно, что алгоритм LRP на данном модельном примере сработал наилучшим образом.

Необходимо отметить, что в решающем правиле распознавания (в узлах дерева) оказываются задействованы не все признаки.

Таким образом, одновременно с построением решающего правила решается задача выбора подсистемы информативных признаков.

На рис.2 приведено дерево решений, построенное по результатам обучения одного из 9 экспериментов, для данного модельного примера.

§ 4. Результаты сравнения LRP с четырьмя алгоритмами распознавания образов на прикладных задачах

Для эксперимента сравнения были выбраны те же четыре алгоритма, что и в § 3. Качество работы каждого алгоритма определялось оценкой вероятности ошибочной классификации на " скользящем экзамене" – $\bar{P}^c_{\alpha}$. Сравнение пяти алгоритмов проводилось для трех реальных задач: распознавания больных с глубоким наркозом и больных с поверхностным наркозом при операции на сердце (Институт патологии кровообращения МЗ РСФСР) по анализу концентрации углекислого газа в артериальной крови (первая задача), в капиллярной крови (вторая задача), в венозной крови (третья задача). Данные для третьей задачи представлены в приложении 3. Число признаков в задачах – 6 (все признаки количественные), объем выборки – I4 для больных с глубоким наркозом и I2 – с поверхностным наркозом. В таблице представлены результаты сравнения.

Таблица

№ задачи	Вероятность ошибочной классификации				
	$\bar{P}^c_1$	$\bar{P}^c_2$	$\bar{P}^c_3$	$\bar{P}^c_4$	$\bar{P}^c_5$
I	0.39	0,35	0,29	0,6I	0.I3
2	0.36	0.54	0.29	0.46	0.I2
3	0.5	0.54	0.36	0.36	0.II

Необходимо отметить, что при сравнении алгоритмов не учитывались положительные свойства LRP : возможность работать с пропусками в таблицах, с несколькими образами и с разнотипными данными, возможность в вершине дерева рассматривать в качестве условия деления гиперплоскость.

Несмотря на то, что положительные свойства алгоритма в эксперименте сравнения не были учтены, алгоритм LRP показал наилучшее качество решения по сравнению с четырьмя известными алгоритмами. Этот факт объясняется следующим: по сравнению с первым и вторым алгоритмами LRP использует более слабое предположение о виде законов распределения вероятностей; по сравнению с третьим алгоритмом LRP дал лучший результат, так как логические решающие правила являются статистически более устойчивыми [7]; по сравнению с четвертым алгоритмом, также основанным на логических решающих функциях, LRP использует более сложные предикаты (в частности, "двухсторонние интервалы") и имеет более сильный критерий [I].

Литература

I. ЛБОВ Г.С., СТАРЦЕВА Н.Г. Алгоритм многоклассового распознавания, основанный на логических решающих функциях.- Настоящий сб., с. 3-I0.

2. ВАПНИК В.Н., ЧЕРВОНЕНКИС А.Я. Теория распознавания образов. - М.: Наука, I974. - 4I5 с.

3. ФУКУНАГА Е. Введение в статистическую теорию распознавания образов. - М.: Наука, I979. - 367 с.

4. ЧЕПКАШИН Н.Т. Некоторые непараметрические алгоритмы распознавания образов балльной размерности. - В кн.: Математическая статистика и ее приложение. Томск, I979, с.I56-I62.

5. МАНОХИН А.Н. Методы распознавания образов, основанные на логических решающих функциях.- В кн.: Эмпирическое предсказание и распознавание образов (Вычислительные системы, вып.67). Новосибирск, I976, с.42-53.

6. МАНОХИН А.Н., СТАРЦЕВА Н.Г. Экспериментальное сравнение 4-х алгоритмов распознавания образов.- В кн.: Обнаружение эмпирических закономерностей с помощью ЭВМ (Вычислительные системы, вып. I02). Новосибирск, I985, с.I27-I32.

7. ЛБОВ Г.С. Методы обработки разнотипных экспериментальных данных. - Новосибирск: Наука, I98I. - I60 с.

Поступила в ред.-изд.отд.
24 мая I985 года

Обращение к программе LRP и описание параметров

Обращение к программе: CALL LRP(X,M,N,Y,M1,K,B,IR,KWT,NTP, NRE,LM,KBET,KGAM,KALF,EPS,MOB,LD,IPRINT,NJF,NWX, APF, OSS,NF, F RA2,SI,E,KOB,C,IR2,NOB,ABE,OSE,MR1,RA3,M2,X1).

Входные параметры:

X - обучающая матрица размерности M x N,

M - количество реализаций обучающей матрицы,

N - количество признаков,

Y - контрольная матрица размерности M1 x N,

M1 - количество реализаций контрольной матрицы,

K - количество образов,

M2 - постоянная, если есть "скользящий экзамен" M2 = M - 1 ; если нет, то M2 = 1 ,

B - код пропуска (кодируется максимальным положительным значением, не встречающимся в таблице),

IR - параметр режима, определяет вариант критерия,

KWT - максимально-возможное количество вершин (целое, нечетное),

NTP - массив размерности N x 2, первый столбец определяет тип признака: 0 - признак наименования, I - количественный, 2 - дискретно-количественный, 3 - балльный; второй столбец рабочий,

NRE - параметр режима работы, определяющий вид выбора условия в вершине,

LM - максимальное количество имен в одном признаке,

KBET, KGAM, KALF, EPS, LD - параметры, регулирующие качество решающего правила,

MOB - массив размерности M x 3, первый столбец - вектор принадлежности реализаций к образу, второй столбец - рабочий, третий - выходной - указывает номер вершины, в которую попадает объект обучающей выборки,

IPRINT - параметр, регулирующий печать: IPRINT = 0 - печати нет, IPRINT = 1 - есть вся печать, IPRINT = 2 - печать только результатов обучения,

IR2 - параметр режима.

Выходные параметры:

OSS - количество ошибок на обучении,

OSE - количество ошибок на "скользящем экзамене",

NJF - массив размерности 4 x KWT , первый столбец - тип высказывания: 0 - набор имен, I - интервал, 2 - набор баллов или дискретов, 3 - гиперплоскость; второй столбец - номер признака, по которому строится высказывание; третий и четвертый - номера признаков для гиперплоскости,

NWX - массив размерности K x KWT - определяет количество объектов каждого образа в вершине,

APF - массив размерности 4 x KWT пороговых значений, перечень имен для высказываний типа 0 (количество имен в наборе не более трех), если один из элементов массива равен B, то для соответствующей вершины это символ конца перечня имен или интервал для высказываний типа I,2; третий и четвертый элементы строки равны 0 или коэффициенты гиперплоскости для 3,

NF - массив размерности KWT x 4 , первый столбец указывает номер вершины, в которую следует идти, если высказывание истинно, второй столбец определяет номер вершины, в которую следует идти, если высказывание ложно, третий столбец указывает номер текущей вершины, четвертый столбец - тип вершины (для конечной вершины значение равно 2),

F - вектор размерности KWT значений критерия.

NOB - вектор размерности M1 - указывает номер предсказанного образа,

ABE - массив условных вероятностей размерности M1 x K.

Рабочие массивы:

RA2 - размерности M x 2
SI - размерности LM x 2
E - размерности 2 x N
KOB - размерности K x 3
C - размерности N x 2
MR1 - размерности M2 x 3
RA3 - размерности M2 x 2
X1 - размерности M2 x N

Результаты работы программы

Исходные данные:

Количество объектов M = 20
Количество признаков N = 6
Параметр KBET = 1
Параметр KGAM = 2
Параматр KALF = 1
Параметр ERS = 0,0
Количество классов K = 3
Количество ветвей KWT = 7
Номер критерия IR = I
Длина внешнего пути есть LD = I
Пропуск B = 9.0

Распределение объектов по образам в каждой вершине:

Вершина номер I
6, 6, 8

Вершина номер 2
0, 0, 6

Вершина номер 3
6, 6, 2

Вершина номер 4
0, 6, 2

Вершина номер 5
6, 0, 0

Вершина номер 6
0, 6, 0

Вершина номер 7
0, 0, 2

Результат распознавания:

Количество объектов на контроле M1 = 4
Объект номер I принадлежит образу 3
Условные вероятности принадлежности к образам
0.0 0.0 I.000

Объект номер 2 принадлежит образу 3
Условные вероятности
0,0 0.0 I.000
Объект номер 3 принадлежит образу I
Условные вероятности
0.4286 0.4286 I.I429
Объект номер 4 принадлежит образу 3
Условные вероятности
0.0 0.0 I.000

Таблица I

Результат обучения

K5	Решающее правило					
	NF(K5,3)	NJF(1,K5)	NJF(2,K5)	NJF(3,K5)	NJF(4,K5)	APF(1,K5)
I	I	0	I	0	0	3.0000
2	2	0	0	0	0	0.0
3	3	3	2	3	5	-6.626I
4	6	0	0	0	0	0.0000
5	7	0	0	0	0	0.0
6	4	2	4	0	0	3.00
7	5	0	0	0	0	0.0

K5	Решающее правило					
	APF(2,K5)	APF(3,K5)	APF(4,K5)	NF(K5,1)	NF(K5,2)	NF(K5,4)
I	6.0000	9.0	0.0	2	3	I
2	0.0	0.0	0.0	0	0	2
3	7.0922	-I.4888	-I.6850	4	5	I
4	0.0000	0.0	0.0	0	0	2
5	0.0	0.0	0.0	0	0	2
6	4.00	0.0	0.0	6	7	I
7	0.0	0.0	0.0	0	0	2

Здесь: K5 - номер, NF(K5,3) - номер исходящей вершины, NF(K5,1) - номер вершины, в которую надо идти, если высказывание истинно, NF(K5,2) - номер вершины, в которую надо идти, если высказывание ложно, NF(K5,4) - указывает конечность вершины, NJF(1,K5) - индекс высказывания, NJF(J,K5), $J = \overline{2,4}$,- номера признаков, по которым строится высказывание, APF(J,K5), $J = \overline{I,4}$,- пороги. Количество ошибок при обучении OSS = 0.0.

Таблица 2

№ объекта	№ признака					
	I	2	3	4	5	6
I	5.0	6.0	5.6	6.0	5.8	0.0
2	0.3	0.3	5.7	6.0	6.I	3.7
3	0.5	5.6	5.0	I.8	I4.0	6.8
4	0.0	I2.I	I7.5	0.0	I7.0	I3.0
5	3.9	0.0	4.7	0.0	I7.2	9.I
6	7.4	5.4	7.5	0.0	22.2	I8.0
7	6.8	I0.0	0.0	0.0	0.0	8.5
8	8.0	II.2	I0.I	I4.8	I3.3	4.4
9	4.8	6.3	0.0	7.8	I8.8	I.6
I0	0.0	0.0	3.8	3.9	7.6	3.I
II	II.0	0.0	0.0	0.0	22.I	II.7
I2	8.0	II.6	0.0	0.0	I0.0	7.5
I3	7.5	0.0	3.5	5.0	0.0	5.0
I4	0.0	0.I	8.2	0.0	8.2	4.8
I5	0.0	5.8	5.6	6.0	I9.9	I2.2
I6	0.0	2.2	0.0	II.8	20.3	I0.9
I7	2.2	0.5	3.0	0.0	I5.0	I0.0
I8	0.0	2.9	0.0	8.0	I6.4	I2.8
I9	0.5	0.0	0.0	5.6	I4.9	7.5
20	0.0	6.0	I0.0	0.0	I4.8	I3.2
2I	0.0	0.0	0.0	0.0	I5.0	I0.8
22	6.I	0.I	0.0	7.5	4.8	2.8
23	0.0	0.0	7.6	9.0	I8.2	II.5
24	8.8	0.0	II.I	0.0	22.7	9.3
25	0.0	7.2	7.8	0.0	0.0	I2.8
26	9.5	II.5	0.0	0.0	I3.6	5.5
27	4.4	0.0	0.0	9.8	I5.3	7.7
28	0.0	6.2	9.5	6.2	I8.0	9.5