Применение метода неравномерных покрытий для решения задачи поиска максимума информативности предиката

А.Ю.Горчаков

***Аннотация*—В данной работе рассматривается решение задачи поиска максимума информативности предиката методом неравномерных покрытий. В статье приведен сравнительный анализ метода неравномерных покрытий с «жадным» алгоритмом и методом полного перебора на примере конкретной задачи.**

***Ключевые слова*—задача глобальной оптимизации, методы поиска информативных закономерностей, задача бинарной классификации.**

I.Введение

Задача бинарной классификации формулируется следующим образом. Пусть задано множество объектов X, множество меток Y = {0,1}, и существует целевая функция $y^*: X \to Y$, значения которой $y_i = y^*(x_i)$, известны только на конечном множестве объектов $X_1, \dots, X_n \in X$. Пары «объект-класс» (X_i, y_i) называются прецедентами. Совокупность пар $(X_i, y_i)_{i=1}^n$ называется обучающей выборкой. Задача бинарной классификации заключается в том, чтобы по обучающей выборке научиться восстанавливать зависимость y^*, то есть построить решающую функцию $X \to Y$, которая бы приближала целевую функцию, причем не только на объектах обучающей выборки, но и на всем множестве X.

В случае если данные $X \in R$, некоторые из предлагаемых методов решения задачи [1],[2] предполагают бинаризацию этих данных.

Пусть $\varphi(x)$ некоторый предикат, определенный на множестве объектов X, выделяет достаточно много объектов одного класса *C*, и практически не выделяет объекты другого класса. Введем обозначения:

P – число объектов класса *C* в выборке

p – из них число объектов, для которых выполняется условие $\varphi(x) = 1$

N – число объектов не принадлежащих классу *C в* выборке

n – из них число объектов, для которых выполняется условие $\varphi(x) = 1$

*Работа выполнена при поддержке РФФИ, проект 16-07-00458

А.Ю. Горчаков – старший научный сотрудник Вычислительного центра им. А.А. Дородницина Федерального исследовательского центра «Информатика и управление» Российской академии наук. andrgor12@gmail.com

Информативность предиката $\varphi(x)$ относительно класса $C \in Y$ по выборке $X^l = (X_i, y_i)_{i=1}^l$будем рассчитывать через статистическое определение информативности [2],[5]:

$$I_c(\varphi, X^l) = -ln\frac{C_P^p C_N^n}{C_{P+N}^{p+n}}, \text{где } 0 \le \text{p} \le \text{P}, 0 \le \text{n} \le \text{N}, \quad (1.1)$$

где $C_m^k = \frac{m!}{k!(m-k)!}$ - биномиальные коэффициенты, $0 \le \text{k} \le \text{m}$

Пусть $f: X \to R$ – числовой признак. Зонами значений признака *f*, будем называть предикаты вида:

$$\varphi(x) = [d \le f(x) \le d'], \qquad d < d' \qquad (1.2)$$

Требуется найти такие d и d' , что $I_c(\varphi, X^l) \to max$.

Возьмем для примера выборку из примерно 100000 прецедентов, где $X_i \in [0,1]$, а множество меток Y = {0,1}.

График зависимости функции $I_c(\varphi, X^l)$ от d' , при различных d выглядит следующим образом (см. рис.1):

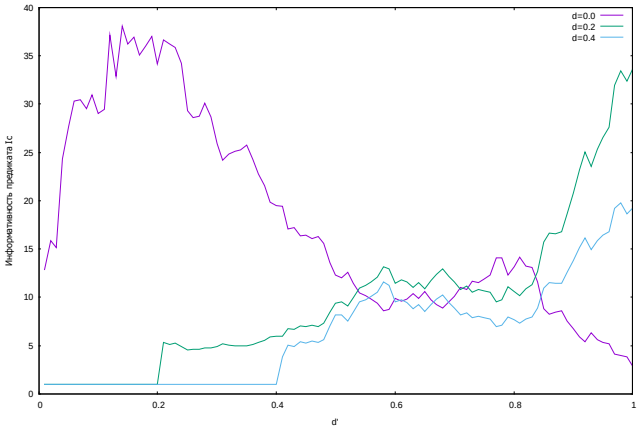

Рис.1

II.Описание алгоритмов.

1.Жадный алгоритм слияния зон [1],[2] – возьмем пороги вида

$$d_i = \frac{f(i) + f(i+1)}{2}, f^{(i)} \ne f^{(i+1)}, i = 1, \dots, l-1 \quad (1.3)$$

где $f^{(1)} \le \dots \le f^{(l)}$- последовательность значений признака f на объектах выборки $f(x_1), \dots, f(x_l)$ упорядоченная по возрастанию. Причем d_i подбираются таким образом, что они проходят между всеми парами точек x_{i-1}, x_i ровно одна из которых принадлежит классу *C*.

Таким образом начальное разбиение состоит их чередующихся зон «только *C* – только не *C*». Далее

зоны укрупняются путем слияния троек соседних зон. Зоны сливаются до тех пор, пока информативность некоторой слитой зоны превышает информативность исходных зон, либо пока не будет получено заданное количество зон r.

Каждый раз выбирается та тройка, при слиянии которой достигается максимальный выигрыш информативности.

2. Метод полного перебора – возьмем пороги вида (1.3) и вычислим значения функции для всех значений $i = 1, \dots, l-2, j = i+1, \dots, l-1$

3.Метод перебора по равномерной сетке –простейший из методов оптимизации действительно-значных функций. Суть метода – разобьем d и d' на n равных частей:

$d_i = \frac{i}{n}, i = 0, \dots, n$ $d'_j = \frac{j}{n}, j = 0, \dots, n$ и вычислим значения функции $I_c(\varphi, X^l)$ в точках $d_i, d'_j\ i = 0, \dots, n, j = 0, \dots, n$, i<j, далее путем сравнения найдем точку в которой функция принимает максимальное значение.

4.Метод неравномерных покрытий (незначительная модификация метода, приведенного в [3],[4]) –

Предположим, что функция $f(x)$ удовлетворяет условию Липшица, то есть для любых x_1 и x_2 существует число $L < 0$ такое, что

$$|f(x_1) - f(x_2)| \le L\|x_1 - x_2\|, \|z\| = \left[\sum_{i=1}^{n} \left(z^{(i)}\right)^2\right]^{\frac{1}{2}} \quad (1.4)$$

и известны ее значения в точках $x_1, x_2, \dots x_k$ из (1.4) следует

$$f(x_k) - L\|x - x_k\| \le f(x) \le f(x_k) + L\|x - x_k\|. \quad (1.5)$$

Определим величину

$$F_k = max[f(x_1), f(x_2), \dots, f(x_k)] \quad (1.8)$$

Найдем множество Δ_k такое, что на Δ_k имеет место

$$f(x) \le F_k + \varepsilon \quad (1.7)$$

Условие (1.7) выполнено для всех x, удовлетворяющих хотя бы одному из k условий

$$f(x_j) + L\|x - x_j\| \le F_k + \varepsilon, j = 1,2, \dots, k. \quad (1.8)$$

При каждом фиксированном j значения x удовлетворяющие (1.8), заполняют n-мерный шар V_j, границей которого является сфера

$$\|x - x_j\| = (F_k - f(x_j) + \varepsilon)/L = R_j \quad (1.9)$$

с центром в точке x_j и с радиусом, равным R_j.

Центры шаров с наименьшими радиусами $R_{min} = \varepsilon/L$ располагаются в тех точках x_j , где $f(x_j) = F_k$. Шар (1.8) и сферу (1.9) будем в дальнейшем обозначать одной буквой V_j.

Величина F_k является решением задачи об отыскании глобального максимума функции *f(x)* на множестве $\Delta_k = \bigcup_{j=1}^{k} V_j$, так как максимальное значение функции *f*, удовлетворяющей (1.7) не превосходит на множестве Δ_k более чем на ε величину F_k.

Если для некоторой последовательности точек $x_1, x_2, \dots x_k$получено Δ_k покрывающее допустимое множество, то тогда F_k есть решение исходной задачи. Способов получения последовательностей таких точек может быть множество. Один из них – разбиваем множество на n-мерные кубы равного размера (по аналогии с методом перебора по равномерной сетке), а далее считаем куб покрытым, если он целиком содержится в одной из n-мерных сфер $V_1, V_2, \dots, V_k$.

Сначала сравним результаты работы 2-х алгоритмов

1. Метод полного перебора
2. «Жадный» алгоритм слияния зон.

Метод	Кол-во вычислений I_c	Максимум I_c	d	d'
Полного перебора	502502	39.287	0.187	0.977
«Жадный» алгоритм	1351	17.901	0.031	0.047

Рис.2

Как видно из рис.2 метод полного перебора находит глобальный максимум, но при этом требует большего количества вычислений. «Жадный» алгоритм останавливает свою работу в локальном максимуме, далеком от глобального.

Далее посмотрим результаты работы метода поиска по равномерной сетке с различными значениями n.

Значение n	Кол-во вычислений I_c	Максимум I_c	d	d'
11	55	34.172	0.0	0.2
51	1275	38.065	0.0	0.14
101	5050	38.065	0.0	0.14
501	125250	38.785	0.0	0.156
1001	500500	39.287	0.187	0.975

Рис.3

Для более корректного сравнения модифицируем алгоритм слияния зон. Разобьем интервал d на n равных подынтервалов и вычислим математическое ожидание y^*, на каждом из них и на всем интервале

$$M = \sum_{i=1}^{l} y_i, \quad (1.10)$$

$$M_j = \sum_{i=1}^{m} y_i, \text{ где } d_j \le f(i) < d_{j+1} \quad (1.11)$$

Введем новый класс C', так что подынтервал принадлежит классу C' если $M_j > M$, и не принадлежит если $M_j \le M$. На вход алгоритма «жадного» слияния зон подаем подынтервалы и новый класс C'.

Значение n	Кол-во вычислений I_c	Максимум I_c	d	d'
11	6	25.860	0.0	0.3
51	50	34.204	0.0	0.24
101	139	36.214	0.0	0.15
501	576	37.757	0.0	0.158
1001	798	35.582	0.0	0.121

Рис.4

Приведем более подробное описание работы алгоритма:
n=11
Шаг1: отрезок [0,1] разбивается на 10 равных подынтервалов. Вычисляется математическое ожидание

$M = 0{,}505$ и M_1= 0,473; M_2=0.490; M_3=0.507; M_4=0.512; M_5=0.518; M_6=0.512; M_7=0.506; M_8=0.502; M_9=0.522; M_{10}=0.515

Шаг2: по формуле (1.3) подбираются пороги d_1 =0.0; $d_2 = 0.3$; $d_3 = 0.8$; $d_4 = 0.9$; $d_5 = 1.0$

Шаг3: находим тройку зон [0.3;0.8] [0.8;0.9] [0.9;1.0], при слиянии которой достигается максимальный выигрыш информативности и сливаем их в одну зону.

Шаг4: в результате слияния зон на шаге 3 осталось 2 зоны [0.0;0.3] [0.3;1.0], из них выбираем зону с максимальной информативностью – [0.0;0.3].

n=51

Шаг1: отрезок [0,1] разбивается на 50 равных подынтервалов. Вычисляется математическое ожидание

Шаг2: по формуле (1.3) подбираются пороги d_1 =0.0; $d_2 = 0.2$; $d_3 = 0.22$; $d_4 = 0.24$; $d_5 = 0.28$; $d_6 = 0.3$; $d_7 = 0.34$; $d_8 = 0.36$; $d_9 = 0.52$; $d_{10} = 0.54$; $d_{11} = 0.6$; $d_{12} = 0.62$; $d_{13} = 0.64$; $d_{14} = 0.66$; $d_{15} = 0.7$; $d_{16} = 0.8$; $d_{17} = 0.94$; $d_{18} = 0.96$; $d_{19} = 1.0$

Шаг3(1): сливаем тройку зон [0.34;0.36] [0.36;0.52] [0.52;0.54].

Шаг3(2): сливаем тройку зон [0.8;0.94] [0.94;0.96] [0.96;1.0].

Шаг3(3): сливаем тройку зон [0.28;0.3] [0.3;0.34] [0.34;0.54].

Шаг3(4): сливаем тройку зон [0.24;0.28] [0.28;0.54] [0.54;0.6].

Шаг3(5): сливаем тройку зон [0.62;0.64] [0.64;0.66] [0.66;0.7].

Шаг3(6): сливаем тройку зон [0.6;0.62] [0.62;0.7] [0.7;0.8].

Шаг3(7): сливаем тройку зон [0.24;0.6] [0.6;0.8] [0.8;1.0].

Шаг3(8): сливаем тройку зон [0.0;0.2] [0.2;0.22] [0.22;0.24].

Шаг4: в результате слияния зон на шаге 3 осталось 2 зоны [0.0;0.24] [0.24;1.0], из них выбираем зону с максимальной информативностью – [0.0;0.24].

Если сравнивать (рис.3, рис.4) алгоритм перебора по равномерной сетке и «дискретизированный» вариант алгоритма слияния зон, то получается, что алгоритм слияния зон имеет существенно меньшую вычислительную сложность, но при этом значение найденного максимума у него ниже чем у алгоритма перебора по равномерной сетке. Причем, начиная с некоторого *n*, качество работы алгоритма ухудшается.

Теперь посмотрим, как работает алгоритм неравномерных покрытий:

Зададим параметры метода

$\varepsilon = 1.0$ и $L = 100.0$

Значение n	Кол-во вычислений I_c	Максимум I_c	d	d'
11	16	33.562	0.2	1.0
51	13	33.653	0.24	1.0
101	12	34.400	0.22	0.99
501	25	28.606	0.25	0.992

Рис.5

$\varepsilon = 1.0, n = 51$

Значение L	Кол-во вычислений I_c	Максимум I_c	d	d'
100	13	33.653	0.24	1.0
200	45	36.289	0.12	1.0
400	147	38.065	0.0	0.14
800	538	38.065	0.0	0.14

Рис.6

Из рис.5 и рис.6 видно, что для работы алгоритма неравномерных покрытий существенно важна оценка константы Липшица L. При заниженной константе метод неравномерных покрытий пропускает точку, в которой функция принимает максимальное значение, при завышенной – производится излишнее количество вычислений функции.

Далее на рис.7-рис.16 приводится первые 10 шагов работы метода с параметрами $\varepsilon = 1.0, L = 100.0, n = 51$. Легенда: «красный» - текущий найденный максимум, «желтый» - точки в которых вычислялась функция информативности предиката I_c , «зеленый» - покрытие при текущем вычислении функции, «синий» - покрытие при обновлении значения найденного максимума.

Рис.7, рис.8 покрытие осуществляется кубами минимального размера (вписанными в сферу радиуса $R_{min} = \varepsilon/L$);

рис.9 радиус покрытия увеличивается;

рис.10 обновлено значение максимума, производится покрытие в окрестностях ранее вычисленных точек;

рис.11, рис12 аналогично рис.9

рис.13 существенно обновлено значение максимума

рис.14, рис.15, рис.16 заключительные шаги алгоритма – покрытие оставшегося множества.

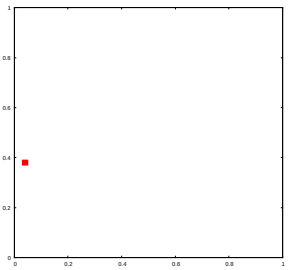

Рис.7

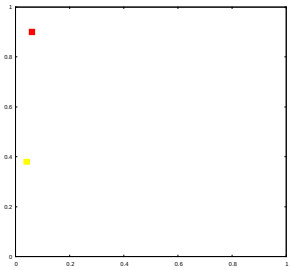

Рис.8

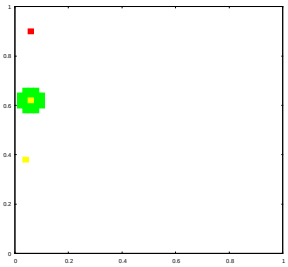

Рис.9

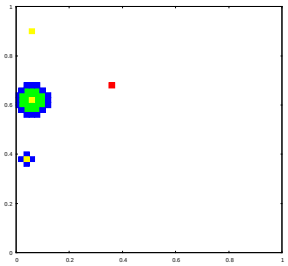

Рис.10

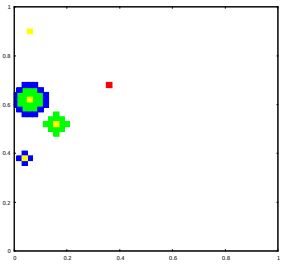

Рис.11

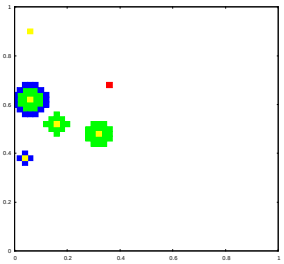

Рис.12

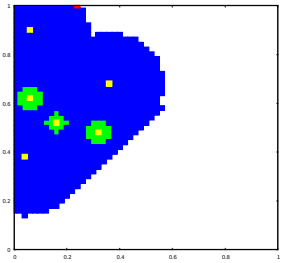

Рис.13

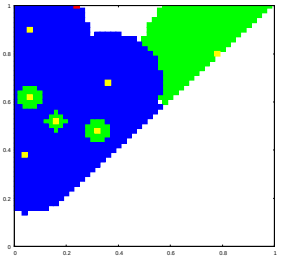

Рис.14

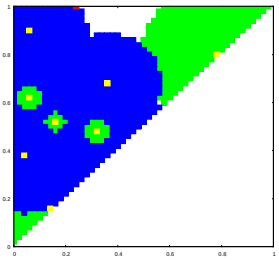

Рис.15

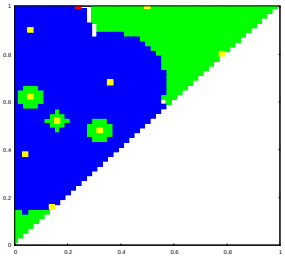

Рис.16

III.Заключение

Сравнительный анализ алгоритмов показал, что метод перебора по равномерной сетке гарантированно находит максимум с заданной точностью, но требует произведения большого количества вычислений. «Жадный» алгоритм слияния зон обходится небольшим количеством вычислений, но нахождение максимума не гарантируется.

Метод неравномерных покрытий, по вычислительной сложности аналогичен алгоритму «жадного» слияния зон и качеству нахождения максимума аналогичен методу перебора по равномерной сетке. Причем, в случае корректной оценки константы Липшица, метод неравномерных покрытий гарантированно находит значение глобального максимума с заданной точностью.

IV.Библиография

[1] Кузьмич Р.И., Гулакова Т.К., Масич И.С. Способы бинаризации разнотипных признаков в задачах классификации //Актуальные проблемы авиации и космонавтики, vol. 6, 2010, pp. 323-325.

[2] Воронцов К. В. Математические методы обучения по прецедентам (теория обучения машин), Москва, 2011.

[3] Евтушенко Ю. Г. Численный метод поиска глобального экстремума функций (перебор на неравномерной сетке) //Журнал вычислительной математики и математической физики, 1971, vol. 6. – pp.1390-1403.

[4] Evtushenko Y., Posypkin M. A deterministic approach to global box-constrained optimization //Optimization Letters, 2013, vol. 4, pp. 819-829.

[5] Dubner P. N. Statistical tests for feature selection in KORA recognition algorithms // Pattern Recognition and Image Analysis, 1994, Vol. 4, no. 4. p. 396.

Application of method nonuniform coverings for maximum information content of predicate search

Andrei Y. Gorchakov

***Abstract*—In this paper, we consider the solution of the maximum information content of predicate search with method nonuniform coverings. The paper presents a comparative analysis of the nonuniform covering method with "greedy" algorithm and the method of exhaustive search for an example problem.**

***Keywords*—the problem of global optimization, search methods informative laws, binary classification problem.**