УДК 004.93

МЕТОД ВЫДЕЛЕНИЯ ИНФОРМАТИВНЫХ ПРИЗНАКОВ В ЗАДАЧЕ РАСПОЗНАВАНИЯ ОБРАЗОВ С УЧИТЕЛЕМ

Колмаков Р.Г.,
научный руководитель д-р техн. наук Медведев А. В.
Сибирский федеральный университет

Введение

На сегодняшний день задача выделения информативных признаков является одной из важнейших задач при построении и настройке систем распознавания образов. Но существующие методы и подходы, например, методы факторного анализа, многомерное шкалирование и др. не всегда подходят для использования на практике и поэтому для решения конкретных задач зачастую применяют эвристические методы.

В качестве альтернативы существующим методам предлагается метод выделения информативных признаков, основанный на анализе коэффициентов размытости. Задача распознавания на основе имеющегося множества прецедентов (образов, класс которых уже известен) называется классификацией с обучением (или с учителем).

Постановка задачи

Пусть X – векторная случайная величина, характеризующая некий объект (образ) в двуальтернативной задаче распознавания (классификации).

$$x_i = (x_i^1, x_i^2, \dots, x_i^{m-1}, x_i^m), i = \overline{1,n}, \quad (1)$$

где m – количество признаков;

n – объем выборки.

Значения признаков формировались функцией random (равномерный закон распределения) в диапазоне [0;3].

Каждому образу x_i, где $i = \overline{1,n}$ ставится в соответствие число y_i.

$$y_i = \begin{cases} 1, \text{если } \forall x_i^j < c \\ -1, \text{иначе} \end{cases}, \quad (2)$$

где y_i – указания учителя о принадлежности x_i к одному из заранее выделенных классов.

$y_i = 1$ (*i-ое* наблюдение относится к первому классу), когда значения каждого из признаков текущего образа не превышают пороговую величину c (величина c задается такой, чтобы количество представителей каждого класса было сопоставимо). Если же значение хотя бы одного из признаков превышает величину c, то образ относится ко второму классу: $y_i = -1$.

Необходимо построить разделяющую поверхность, которая позволит оптимально (с точки зрения минимума ошибок классификации) разбить ранее не расклассифицированные образы на классы.

Непараметрический алгоритм

При построении решающего правила (классификатора) была использована оценка Розенблатта-Парзена:

$$\varphi_k(x) = \sum_{i=1}^{n} y_i \prod_{j=1}^{m} \Phi\left(\frac{x^j - x_i^j}{Cs_j}\right), \quad (3)$$

$\Phi\left(\frac{x^j - x_i^j}{Cs_j}\right)$ – ядро (весовая функция);

Cs_j – коэффициент размытости j-го признака.

Используется треугольное ядро:

$$\Phi\left(\frac{x^j - x_i^j}{Cs_j}\right) = \begin{cases} 1 - \frac{\left|x^j - x_i^j\right|}{Cs_j}, Cs_j > \left|x^j - x_i^j\right| \\ 0, Cs_j \leq \left|x^j - x_i^j\right|, \end{cases} \quad (4)$$

где $i = \overline{1,n};\ j = \overline{1,m};$

x^j – значение j-го признака классифицируемой точки;

x_i^j – значение j-го признака -ой точки обучающей выборки;

Решение о принадлежности точки к одному из заранее выделенных классов (в нашем случае 2 класса) принимается исходя из значения решающего правила:

$\varphi_k > 0$; x – принадлежит первому классу.

$\varphi_k < 0$; x – принадлежит второму классу.

Как видно из формулы 3 для каждого признака будет свой собственный коэффициент размытости. В качестве кандидата на выброс (как неинформативный признак) предлагается рассматривать признак с наибольшим коэффициентом размытости Cs^-.

$$Cs^- = \max Cs_j, \qquad j = \overline{1,m}; \quad (5)$$

Данная методика была предложена доктором технических наук Медведевым А.В. Признак с наибольшим коэффициентом размытости окажет наименьшее влияние на значение $\varphi_k(x)$, т.е. его вклад будет минимальным.

Если принять все вышеперечисленное, то самой важной частью работы данного алгоритма является собственно настройка и оптимизация коэффициентов размытости с точки зрения минимального числа ошибок классификации.

Предлагается использовать глобальный многомерный поиск с небольшим изменением на начальной стадии. Точки на первой итерации не будут браться абсолютно случайно. Т.е будет проведена классификация (при одинаковом коэффициенте размытости для всех признаков) с шагом Δ (рисунок 1).

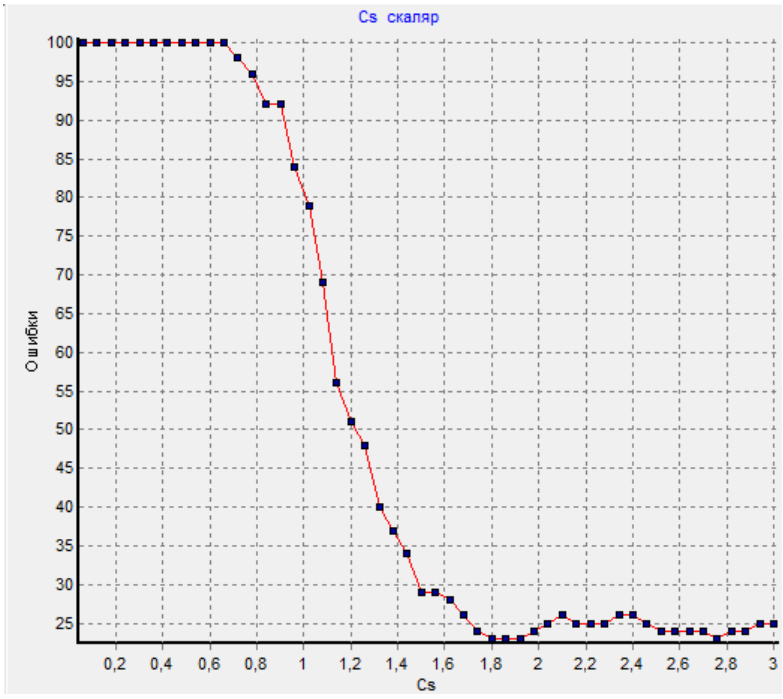

Рисунок 1 – Результаты классификации при скалярном коэффициенте размытости.

Область значений коэффициентов размытости примем аналогичной области значений признаков [0;3]. На рисунке 1 можно видеть результат работы алгоритма

при скалярном коэффициенте размытости (одинаковом для всех признаков). Как видно из рисунка, минимальное значение ошибок классификации было достигнуто при $Cs = 1{,}9$. Это значение и будет использоваться на первой итерации оптимизации коэффициентов размытости как базовое. Это позволит дать «неплохой старт» для поиска оптимальных коэффициентов размытости.

В дальнейшем на каждой итерации формируется k наборов коэффициентов размытости, путем прибавления к базовому набору коэффициентов размытости по каждому признаку случайной добавки:

$$Cs^* = Cs_j \pm \gamma, \quad (6)$$

где $j = \overline{1, m}$;

γ– случайная величина, распределенная по равномерному закону.

Набор, при котором число ошибок классификации было минимальным, берется как базовый на следующих итерациях.

После оптимизации коэффициентов размытости алгоритм производит проверку на неинформативность признака с наибольшим коэффициентом размытости, т.е. проводит классификацию без учета данного признака. Если количество ошибок классификации уменьшилось или не изменилось, то принимается решение об отсеве данного признака и заново запускается алгоритм настройки коэффициентов размытости, но уже без учета данного признака. Так происходит до тех пор, пока признаки не перестанут отсеиваться. Если после очередного этапа работы программы признак не был отсеян, то классификация считается завершенной.

Вычислительные эксперименты

Рассмотрим результат работы программы при объеме выборки 1000, 5 признаках один из которых неинформативен (рисунок 2).

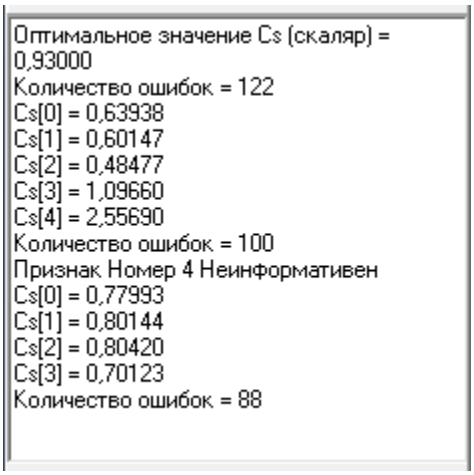

Рисунок 2 – Результаты классификации при n=100, m=5, 4-ый признак неинформативен.

Коэффициент размытости неинформативного признака значительно превышает значения всех остальных (2,55 по отношению к 0,63, 0,60 и т.д.). Отсеяв данный признак, удалось сократить количество ошибок на 12(1,2%).

Далее приведена сводная таблица результатов работы программы. Были рассмотрены задачи классификации максимально приближенные к реальности, т.е. объем выборки не столь велик по отношению к количеству признаков.

N	Представ. Класса (1 - 2)	m	Неинф. признаков	Ошибок Cs скаляр	Ошибок после оптими-зации	Призна-ков отсеяно	Итого ошибок	Процент ошибок (%)	Время работы, с
100	52-48	30	5	49	26	15	16	16	594,906
100	55-45	30	15	42	20	19	19	19	401,000
100	53-47	30	25	51	17	12	13	13	1703,92
100	52-48	50	0	47	22	24	19	19	1161,46
100	52-48	50	25	34	20	33	14	14	824,703
100	41-59	50	45	44	14	20	11	11	870,515
100	51-49	50	49	26	0	48	0	0	605,25
150	76 – 74	80	0	69	63	60	28	18,6	51,547
150	71 – 79	80	40	60	41	34	21	14,0	343,531
150	71-79	150	0	64	58	98	37	24,6	263,406
500	252-248	50	0	228	199	31	173	34,6	1167,83
500	272-228	50	15	214	192	26	170	34,0	780,62
500	254-246	50	25	223	181	31	157	31,4	1369,86
500	227-273	50	40	221	172	34	135	27,0	1068,34
500	245-255	50	45	210	157	36	92	18,4	408,80
1000	477-523	30	0	472	416	14	373	37,3	1587,65
1000	512-488	30	15	420	394	10	376	37,6	420,46
1000	506-494	30	25	309	203	22	176	17,6	231,584
3000	1498-1502	30	0	1372	1346	5	1287	42,9	3968,29
3000	1510-1490	30	25	945	575	25	396	13,2	1724,01

Данный метод выбора признаков позволяет сократить количество ошибок классификации и отсеять наименее неинформативные признаки. Это позволяет увеличить скорость работы классификации при последующем использовании. При наличии неинформативных признаков разработанному алгоритму удается сократить количество ошибок классификации путем настройки коэффициентов размытости по каждому признаку в среднем на 40-55%, после удаления неинформативных признаков еще на 25-50%. При отсутствии неинформативных признаков настройка коэффициентов размытости позволяет избежать 10-35% ошибок, отсев менее информативных признаков 5-10%.