ВЫЧИСЛИТЕЛЬНЫЕ СИСТЕМЫ
Сборник трудов
Института математики СО АН СССР
1967 г. Выпуск 28

ОБ ОПРЕДЕЛЕНИИ ОБЪЕМА ОБУЧАЮЩЕЙ ВЫБОРКИ ЛИНЕЙНОГО КЛАССИФИКАТОРА

Ш.Ю. Раудис
(Вильнюс)

I. Введение

Известно [I], что вероятность ошибки классификатора P_N, обученного выборкой конечного объема N, является случайной величиной. Вероятность ошибки P_N всегда будет не меньше своего минимального значения P_∞, полученного при бесконечном объеме обучающей выборки (ООВ). При ограниченном ООВ вероятность ошибки классификатора (ВОК) может в несколько раз превысить свое минимальное значение [2], поэтому во многих практических задачах опознавания, в частности в устройствах, опознающих речевые команды и приспосабливающихся к каждому диктору, где ООВ – одна из основных характеристик, определение минимального объема обучающей выборки является одной из главных задач.

Данная работа, определяющая ОВВ в случае классификации двух классов объектов, признаки которых распределены нормально с одинаковыми и известными ковариационными матрицами, является продолжением работ Бравермана [3], рассматривающего одномерный случай, и Лбова [I], экспериментально определившего объем представительной выборки в случае единичной ковариационной матрицы.

2. Выбор критерия

Критерий, по которому определяется представительность выборки, зависит от типа решаемой задачи.

При сравнении нескольких классификаторов или систем признаков для того, чтобы выбрать наилучший существенно знать точность определения качества классификации. Поэтому здесь нужно знать характеристики разброса распределения $F(P_N)$: дисперсию, вероятность попадения P_N в интервал заданной длины и т.п. (см., напр. Лбов[I], где в качестве критерия использована длина интервала, в который с данной вероятностью должен попасть экспериментальный результат).

Решая другие задачи (например, определение параметров (обучения) классификатора, годности системы признаков и т.д.), существенно знать характеристики, указывающие местоположение случайной величины P_N , отношение математического ожидания или моды распределения $F(P_N)$ к соответствующим характеристикам, полученным при бесконечном N , вероятности, что P_N не превысит свое минимальное значение на некоторую величину и т.п.

Для определения "достаточности обучения" классификатора важно знать, насколько близка вероятность ошибки обученного классификатора к минимальной, которая может быть получена при увеличении объема обучающей выборки до бесконечности, поэтому в качестве критерия представительности обучающей выборки используем отношение математического ожидания вероятности ошибки при конечном объеме выборки MP_N с вероятностью ошибки P_∞ при бесконечном N, т.е.

$$\varkappa = MP_N / P_\infty . \qquad (I)$$

С помощью критерия $\varkappa$ можно указать доверительный интервал случайной величины P_N . Ширина интервала будет зависеть от имеющихся сведений о характере закона распределения $F(P_N)$.

В случае отсутствия сведений, минимаксным решением будет выбор $F(P_N)$, распределенного, как две δ -функции: одна в точке P_∞ , другая - у конца доверительного интервала. Тогда доверительный интервал будет $(P_\infty, (1+(\varkappa-1)\beta)P_\infty)$,

где $\beta = 1/(1-\alpha)$

α - доверительный уровень.

Экспериментальные результаты [I] показывают, что в случае нормального распределения признаков, распределение $F(P_N)$ близко к нецентральному χ^2 распределению. Задавшись наиболее неблагоприятным, в смысле ширины доверительного интервала , из

данного класса распределением. центральным χ^2 распределением с одной степенью свободы, получаем приближенное значение доверительного интервала

$$(P_\infty, (1+(\varkappa-1)\beta)P_\infty),$$

где $\beta = \chi_1^2$

χ_1^2 и $(1-\alpha)$ - процентное значение величины χ^2 с I степенью свободы.

Следует заметить, что более точное описание распределения $F(P_N)$ может еще сузить доверительный интервал. Об этом свидетельствует и анализ экспериментального материала Лбова [I] , где в 99% случаев β колебалось в диапазоне от I,5 до 3.

3. Объем обучающей выборки при известной ковариационной матрице

Вероятность ошибки классификатора, её прирост, тем самым и объем обучающей выборки будет зависеть от типа решающего правила, имеющихся априорных сведений о законах распределений признаков,и от самих распределений. При выборе зависимостей, определяющих ООВ, будем считать характеристики распределений известными. При практическом определении ОВВ будет требоваться знание оценки нужных характеристик распределений (P_∞, η)

В рассматриваемом нами случае, когда обе классифицируемые совокупности распределены нормально с одинаковыми и известными ковариационными матрицами, оптимальным решающим правилом будет линейная дискриминантная функция [4,5] :

$$W = [X - \tfrac{1}{2}(m^{(1)} + m^{(2)})]' \Sigma^{-1} (m^{(1)} - m^{(2)}),$$

где X - классифицируемый p -мерный вектор признаков,

$m^{(1)}, m^{(2)}, \Sigma$ - выборочные средние и ковариационная матрица распределений соответственно.

Так как вероятности ошибок обоих родов равны, то, рассматривая случай, когда X принадлежит I классу, имеем

$$MP_N = \iint\limits_{\Gamma\; W<0} f_1(x) f(m^{(1)}, m^{(2)})\, dx\, dm^{(1)} dm^{(2)},$$

где $f_1(x)$ – плотность распределения признаков первой совокупности в области Γ их существования;

$f(m^{(1)}, m^{(2)})$ – плотность распределения выборочных средних $m^{(1)}$ и $m^{(2)}$, при этом предполагается, что цены ошибок обоих родов и вероятности появления обоих классов равны между собой.

Известно [6], что при вышеописанных условиях W распределена, как

$$w = \chi_1^2 - \chi_2^2,$$

где χ_1^2 и χ_2^2 – независимые случайные величины, имеющие нецентральное χ – квадрат распределение с p степенями свободы и параметрами λ_1 и λ_2, соответственно:

$$\lambda_i = N\delta^2/4\ [1+(-1)^i(1+2N)^{-1/2}];$$

$$\delta^2 = (\mu^{(1)}-\mu^{(2)})^T \Sigma^{-1} (\mu^{(1)}-\mu^{(2)}); \qquad (2).$$

$\mu^{(1)}$ и $\mu^{(2)}$ – средние распределений.

Тогда

$$MP_N = P_{rob}(w<0) = P_{rob}(\chi_1^2/\chi_2^2 < 1).$$

Точное значение MP_N можно подсчитать по методике [7], однако для практических целей с довольно высокой точностью MP_N можно найти, аппроксимируя распределения $\sqrt[3]{\chi_1^2}$ и $\sqrt[3]{\chi_2^2}$ нормальными [6,8].

$$MP_N = \Phi\{[9(r_1+r_2)+2(b_2 r_2^{-2}-b_1 r_1^{(-2)})]\,[18(b_1 r_1^{-1}+b_2 r_2^{-1})]^{-1/2}\}, \qquad (3)$$

где

$$r_i = (p+\lambda i)^{1/3};$$

$$b_i = 2 - p(p+\lambda_i)^{-1};$$

$$\Phi(c) = (2\pi)^{-1/2}\int_{-\infty}^{c} exp(-v^2/2)\,dv.$$

Таблица I

Объем представительной выборки по критерию æ в случае априори известной ковариационной матрицы

p	æ \ P_∞	0,000I	0,0005	0,00I	0,005	0,0I	0,05	0,I	0,2
I	I,2	25 (20)	I8 (I6)	I6(I5)	I0(I0)	7(8)	4(5)	2(3)	2(2)
	I,5	II (9)	8 (7)	7(7)	4(5)	3(4)	2(2)	I(I)	I(I)
	2,0	6 (5)	5 (4)	4(4)	3(3)	2(2)	I(I)	I(I)	I(I)
3	I,2	26	22	20	I4	II	7	5	5
	I,5	I2	9	8	6	5	3	3	2
	2,0	7	5	5	4	3	2	2	I
5	I,2	3I	24	22	I6	I3	I0	9	9
	I,5	I3	II	9	7	6	5	4	3
	2,0	8	6	5	5	4	3	3	I
8	I,2	35	28	27	20	I8	I4	I3	I3
	I,5	I5	I2	II	I0	9	7	6	5
	2,0	9	7	6	6	6	4	4	I
I2	I,2	4I	35	34	27	24	20	20	20
	I,5	I8	I5	I4	I2	II	9	8	7
	2,0	I0	8	8	8	6	5	4	2
20	I,2	52	45	43	39	36	28	32	34
	I,5	23	20	I9	I6	I5	I3	I2	I0
	2,0	I3	I2	II	II	9	7	7	2
50	I,2	I00	90	89	84	8I	80	80	85
	I,5	42	40	39	36	35	33	30	24
	2,0	24	23	22	2I	I9	I6	I5	5
I00	I,2	I60	I60	I60	I60	I60	I60	I60	I70
	I,5	74	70	69	67	65	68	60	45
	2,0	43	42	37	35	34	3I	30	8

Расстояние между классами δ можно найти из соотношения:

$$P_\infty = \Phi(-\delta/2) \qquad (4)$$

При помощи уравнений (I) - (4) для каждого данного P_∞ ρ и $\varkappa$ можно найти объем представительной выборки.

В табл. I приведен объем представительной обучающей выборки по критерию $\varkappa$ при разных значениях P_∞, ρ и $\varkappa$. В скобках даны точные значения $\mathcal{N}$ (при ρ = I). При увеличении ρ и $\mathcal{N}$ распределение $\sqrt[3]{\chi^2}$ быстро сходится к нормальному [8] и точность определения $\mathcal{N}$ увеличивается.

4. Объем обучающей выборки при неточном определении ковариационной матрицы

В предыдущей главе получена зависимость $\mathcal{N}$ от P_∞, ρ и $\varkappa$ в случае, когда ковариационная матрица известна точно. На практике ковариационная матрица бывает известна неточно, либо принимается диагональный или единичный [9]. Так как Σ неизвестна, то точно определить $\mathcal{N}$ принципиально невозможно. Покажем, что и в этом случае вышеполученные результаты могут быть использованы при введении поправки, учитывающей неточность определения Σ.

Пусть истинное значение Σ оценено как A, тогда

$$w = [X - 1/2(m^{(1)} + m^{(2)})]' A^{-1}(m^{(1)} - m^{(2)})$$

Известно, что если A и Σ - положительно определенные матрицы, то возможна такая линейная трансформация $Y = FX$, что $FAF' = D$ и $F\Sigma F' = 1$, где D - диагональная, а I - единичная матрицы.

Тогда

$$w = (Z + t/2)' D^{-1}(V + t),$$

где $t = F(m^{(1)} - m^{(2)})$, $t' = (t_1, t_2, \ldots, t_p)$;

Z распределено, $\mathcal{N}(0,\ (1 + 1/(2\mathcal{N}))\cdot I)$;
V распределено, $\mathcal{N}(0,\ 2/\mathcal{N}\cdot I)$,

Z и V независимы.

При увеличении ρ, $\mathcal{N}$ и t_i распределение w стремится к нормальному и математическое ожидание вероятности ошибки будет

$$MP_{\mathcal{N}} = \Phi[-Mw \cdot (Dw)^{-1/2}],$$

где

$$Mw = 1/2 \sum_{i=1}^{p} d_i t_i^2;$$

$$Dw = \sum_{i=1}^{p} \{d_i^2 t_i^2 + 1/\mathcal{N}(d_i^2 t_i^2 + 2d_i^2) + 1/\mathcal{N}^2 d_i^2\},$$

d_i - элемент диагональной матрицы D^{-1}

Обозначив

$$(\sum_{i=1}^{p} d_i t_i^2)^2 (\sum_{i=1}^{p} d_i^2 t_i^2)^{-1} = \delta^{*2} \ (\delta^* \le \delta), \quad \text{и}$$

$$0 < \gamma = \rho(\sum_{i=1}^{p} d_i^2 t_i^2)^2 (\sum_{i=1}^{p} d_i^2)^{-1} (\sum_{i=1}^{p} d_i t_i^2)^{-2} < \rho,$$

имеем

$$MP_{\mathcal{N}} = \Phi\{-\delta^*/2[1 + 1/\mathcal{N}(1 + 2\rho/\gamma\delta^{*2}) + 1/\mathcal{N}^2 \rho/\gamma\delta^{*2}]\}^{-1/2} \quad (5)$$

В случае $A = \Sigma$, $\delta^* = \delta$, $\gamma = 1$ и выражение (5) приобретает вид:

$$MP_{\mathcal{N}} = \Phi\{-\delta/2[1 + 1/\mathcal{N}(1 + 2\rho/\delta^2) + 1/\mathcal{N}^2 \rho/\delta^2]^{-\frac{1}{2}}\} \quad (6)$$

Сравнение выражений (5) и (6) показывает, что в случае неточной оценки Σ, объем представительной обучающей выборки $\mathcal{N}^*$ приблизительно можно оценить по приведенным в параграфе 3 формулам или табл. I. При этом полученное значение $\mathcal{N}$ следует умножить на коэффициент несоответствия η: $\mathcal{N}^* = \eta\,\mathcal{N}$, где

$$\eta = [1+2p/(\gamma\delta^{*2})][1+2p/\delta^{*2}]^{-1}$$

$$\gamma = \frac{p\{[F(\mu^{(1)}-\mu^{(2)})]' D^{-2} F(\mu^{(1)}-\mu^{(2)})\}^2}{t \sim D^{-2}\{[F(\mu^{(1)}-\mu^{(2)})]' D^{-1} F(\mu^{(1)}-\mu^{(2)})\}^2};$$

F - матрица, удовлетворяющая условиям

$F\Sigma F' = I$ и $FAF' = D$ (D - диагональная и I -единичная матрицы)

В случае $2p/\delta^{*2} << 1$, $\eta \approx 1$. В других случаях при помощи предварительных экспериментов следует оценить возможные значения γ для данного рода объектов и систем признаков. Как отмечалось, при определении объема представительной выборки на практике также следует оценивать возможное значение P_∞. Чем точнее будет оцениваться P_∞ и η, тем точнее будет определен объем представительной выборки. Выше оценен ООВ в случае независимых реализаций в выборке. На практике, особенно при собирании статистики речевых сигналов, реализации в выборке будут зависимые и представительный ОВВ изменится, поэтому исследование влияния зависимости между отдельными реализациями в выборке требует специального рассмотрения.

5. Выводы

I. Получена зависимость между объемом представительной выборки N, критерием представительности $\varkappa$, количеством признаков и вероятностью ошибки обученного классификатора P_∞ в случае классификации двух классов объектов, признаки которых распределены нормально, с одинаковыми и априори известными ковариационными матрицами.

2. Предлагается методика оценки N в случае неточного определения ковариационной матрицы.

3. Точность оценки N зависит от точности оценки P_∞ и коэффициента несоответствия η.

4. Так как другие критерии представительности выборки могут быть получены из критерия $\varkappa$, то полученные результаты можно использовать и для решения других задач, связанных с представительностью выборки; а именно: сравнения вероятностей ошибок нескольких классификаторов, принятия решения о годности системы признаков и т.д.

Литература

I. Г.С. Лбов. О представительности выборки при выборе эффективной системы признаков. - Вычислительные системы, Новосибирск, 1966, вып. 22, стр. 39-58.

2. Ш.Ю. Раудис и А.Ю. Мотуза. Влияние точности описания исходных данных на средний риск классификаторов разных типов. - Автоматика и вычислительная техника. Труды XУI научно-технической конференции. Издание Каунасского политехн. ин-та (Сдано в печать).

3. Braverman D. "Learning Filters for Optimum Pattern Recognition". IRE Trans.Vol.IT-8.1962 pp.280-285.

4. S.Das Gupta "Optimum classification Rules for Classification into Two Multivariate Normal Populations".Ann.Math.Stat.Vol.36.1965 pp.1174-1184.

5. Т. Андерсон. Введение в многомерный статистический анализ. Изд. ФМ лит. Москва, 1963.

6. John S. "Errors in discrimination" Ann.Math.Stat.Vol. 32.1961.pp.1125-1144.

7. Price R. "Some non-central F-distributions expressed in closed form" Biometrika Vol.51.1965.pp.107-122.

8. Abdel-Aty S.H. "Aproximate formulae for the percentage points and the probability integral of the non-central distribution".Biometrika Vol.41.1954. pp.538-540.

9. И.Я. Левин. Некоторые вопросы теории опознавания образов. Изд. Техническая кибернетика АН СССР, 1964, № 2 стр. 50-55.

Поступила в редакцию
I/X-1967г.