ISSN 0201-8632

УДМУРТСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

Известия
Института математики и информатики

Выпуск 2 (36)

$$\dot{x}(t) = \int_{-r}^{0} dA(t,s)x_t(s),\ t \in \mathbb{R}$$

$$x_t(s) \doteq x(t+s),\ s \in [-r,0]$$

Ижевск 2006

Выпуск содержит труды научной конференции–семинара «Теория управления и математическое моделирование», посвященной 50-летию Ижевского государственного технического университета и 30-летию кафедры прикладной математики и информатики ИжГТУ.

Для специалистов по дифференциальным уравнениям и теории управления.

УДК 519.92

© А. Г. Ицков

ЕМКОСТЬ СЕМЕЙСТВА РЕШАЮЩИХ ПРАВИЛ ПРИ ОБУЧЕНИИ РАСПОЗНАВАНИЮ ОБРАЗОВ

Ключевые слова: распознавание образов, решающее правило, ёмкость, вероятность ошибки.

Abstract. A notion of the capacity of a family of decision rools is inspected. The bounded capacity provides convergence of the frequency of errors to the probability of errors. Some estimates of capacity for the families of linear and non-linear decision rools are given.

В классической постановке задачи обучения распознаванию образов для двух классов требуется из заданного семейства решающих правил (алгоритмов распознавания) $F(x,\alpha)$, где $x \in \mathbb{R}^n$, α – скалярный или векторный параметр из некоторого множества, выбрать правило, минимизирующее средний риск или вероятность ошибки, то есть минимизировать функционал

$$R(\alpha) = \sum_{\omega=0,1} \int_X (\omega - F(x,\alpha))^2 dP(\omega, x), \tag{1}$$

где $\omega = 0,1$ – номер класса, к которому относится $x \in X$, $P(\omega, x)$ – совместное вероятностное распределение на $\{\omega\} \times X$.

Поскольку распределение $P(\omega, x)$ обычно неизвестно, минимизацию (1) заменяют минимизацией функционала эмпирического риска

$$\rho(\alpha) = \frac{1}{l} \sum_{i=1}^{l} (\omega_i - F(x_i,\alpha))^2, \tag{2}$$

который вычисляется по обучающей выборке $x_1, \ldots, x_l$.

Близость точек минимума функционалов (1) и (2) обеспечивается при условии равномерной сходимости $\rho(\alpha)$ к $R(\alpha)$:

$$P\left\{\sup_{\alpha}|\rho(\alpha) - R(\alpha)| > \varepsilon\right\} \underset{l\to\infty}{\longrightarrow} 0. \tag{3}$$

Достаточным условием выполнения (3) является ограниченность функции роста [1] $m(l)$ семейства $\{F(x,\alpha)\}$, определяемой как максимальное число способов разделения l точек на два класса с помощью решающих правил $F(x,\alpha)$. Для $m(l)$ справедлива оценка

$$m(l) \le 1,5\frac{e^h}{h!},$$

где h есть максимальный объем выборки, которую можно поделить на два класса с помощью правил $F(x,\alpha)$ всеми возможными 2^h способами. Число h называется емкостью класса $F(x,\alpha)$. Если же $m(l) \equiv 2^l$, то емкость класса считается бесконечной.

Число h, таким образом, служит мерой разнообразия правил в семействе $F(x,\alpha)$. При этом справедлив следующий результат:

Если на выборке объема l $\rho(\alpha)$ близок к нулю, то с заданной надежностью η вероятность ошибки $R(\alpha)$ на всем пространстве X не превосходит заданной величины $\varepsilon > 0$, где

$$l \simeq \frac{h - \ln\eta}{\varepsilon}. \tag{4}$$

Таким образом, зная емкость класса, можно оценить вероятность получения решающего правила с заданным качеством распознавания. При этом объем обучающей выборки оценивается по формуле (4).

Приведем ряд оценок емкости для некоторых распространенных семейств распознавания.

1) Семейство гиперплоскостей в R^n, проходящих через начало координат.

Здесь $F(x,\alpha)$ определяется уравнением $\sum_{i=1}^{n}\alpha_i x^i = 0$, то есть $\alpha = (\alpha_1,\ldots,\alpha_n)$ – вектор коэффициентов гиперплоскости, $x = (x^1,\ldots,x^n) \in R^n$. В комбинаторной геометрии доказывается, что

$$m(l) = \begin{cases} 2\sum\limits_{i=0}^{n-1} C_{l-1}^{i} & , \quad l \geq n, \\ 2^{l} & , \quad l < n. \end{cases}$$

Отсюда следует, что емкость h равна размерности пространства R^n.

2) Семейство произвольных гиперплоскостей в R^n.

Здесь $F(x,\alpha)$ определяется уравнением $\sum\limits_{i=1}^{n} \alpha_i x^i = \alpha_{i+1}$, этот случай можно свести к предыдущему, погружая данное семейство в пространство R^{n+1}, где $x^{n+1} \equiv 1$. Таким образом, емкость h равна $n+1$.

3) Семейство гиперповерхностей второго порядка.

$F(x,\alpha)$ определяется уравнением $\sum\limits_{i,j}^{n} \alpha_{ij} x^i x^j + \sum\limits_{i=1}^{n} \beta_i x^i = \gamma$. Аналогично предыдущему пункту, погружая данное семейство в семейство гиперплоскостей в спрямляющем пространстве, имеем $h = C_{n+2}^{2} = \frac{(n+1)(n+2)}{2}$. В частности, для гиперсфер $h = n+2$. Для полиномиальных поверхностей порядка r емкость равна C_{n+r}^{r}.

4) Алгоритм "Кора".

Здесь x – бинарный вектор: $x^i \in \{0,1\}, i = 1,2,\ldots,n$. Для каждого из двух классов в обучающей выборке ищется заданное число t конъюнкций $x^i x^j x^k$, для которых нет совпадений в другом классе. Классификация вектора x производится "голосованием" по всем отобранным конъюнкциям.

Поскольку здесь пространство объектов дискретное, то проще оценить число всех решающих правил N:

$$N \leq \left(8C_k^t\right)^2, K = 8C_n^3.$$

Это приводит к очевидной оценке $N \leq n^{6t}$ и, следовательно, емкость h не превосходит $6t \ln n$. Таким образом, для данного семейства объем обучения пропорционален только логарифму размерности пространства.

5) Тестовые алгоритмы.

Тестом бинарной таблицы обучения называется подмножество признаков $\{j_{i_1}, j_{i_2}, \ldots, j_{i_k}\}$, для которого объекты разных классов не совпадают. Тест называется тупиковым тестом, если он является минимальным тестом по включению. Классификация вектора x производится вычислением оценок

$$\Gamma(x) = \sum_{a \in A} \sum_{x_i} \sigma(ax, ax_i)\gamma(x_i)p(a),$$

где A – система тупиковых тестов a, $\gamma_i, p_j \geq 0$ – веса объектов обучения и признаков, σ – функция близости. Объект x зачисляется в тот класс, для которого оценка $\Gamma(x)$ больше.

При вычислении емкости можно показать [2], что данное семейство алгоритмов погружается в семейство гиперплоскостей в пространстве R^{ln}, где l – объем обучения, n – число признаков. Таким образом, емкость семейства тестовых алгоримов не превосходит ln.

Итак, можно констатировать, что для всех рассмотренных семейств алгоритмов существует емкость невысокого порядка, что позволяет успешно решать задачу обучения распознаванию при сравнительно небольшом объеме обучающей выборки.

Список литературы

1. В.Н. Вапник, А.Я.Червоненкис. Теория распознавания образов. М.:Наука, 1974.
2. А.Г. Ицков. О емкости модели распознавания алгоритмов вычисления оценок // Журн. выч. матем. и матем. физ. 1982. 22. №4, с.975-981.