УДК 519.1, ВАК 05.13.18, ГРНТИ 28.29.51

ДАНИЛОВ Г. В., КАЧАН О. В., БОРИСОВА Н. К. ОБ ОЦЕНКЕ ЭФФЕКТИВНОСТИ МЕТОДОВ РАСПОЗНАВАНИЯ ОБРАЗОВ

Об оценке эффективности методов распознавания образов

Г. В. Данилов, О. В. Качан, Н. К. Борисова

Ухтинский государственный технический университет, г. Ухта; Институт психологического консультирования «Новый век», Санкт-Петербург, Россия

Статья посвящена вопросам теоретической оценки эффективности алгоритмов классификации, выводится формула для расчёта гарантированной оценки надёжности распознавания в зависимости от конкретного соотношения объектов A и B в обучающей выборке для случая дихотомии, приведён пример расчётов с применением данной формулы.

Ключевые слова: *классификация, оценка эффективности распознавания, надёжность обучения с учителем, теория вероятности.*

On the assessment of the effectiveness of the methods of pattern recognition

G. V. Danilov, O. V. Kachan, N. K. Borisova

Ukhta state technical University, Ukhta; Institute of psychological counseling "Novyy Vek", Saint Petersburg, Russia

The article is devoted to theoretical evaluation of the efficiency of classification algorithms, deduces a formula for the calculation of the guaranteed estimation of reliability of recognition depending on the specific ratio of the objects A and B in the training set for the case of dichotomy, gives an example of calculations using this formula.

Keywords: *classification, performance evaluation of recognition reliability learning with a teacher, the theory of probability.*

Острая потребность в решении задач, связанных с классификацией объектов по тем или иным признакам, характеризующим их, породила в последние десятилетия поток самых разнообразных приёмов и методов под общим названием «методы распознавания образов», практика применения которых стала значительно опережать их полное теоретическое обоснование. Более того, даже когда точно сформулированы условия применимости того или иного метода, их соблюдение на практике часто не может быть гарантировано.

Что же касается эффективности того или иного метода, то её теоретические оценки на сегодняшний день зачастую отсутствуют и на практике заменяются эмпирическими заключениями, которые, особенно при работе с малыми выборками, абсолютно ненадёжны.

В этих условиях полезно иметь пусть грубую, но универсальную теоретически обоснованную нижнюю оценку эффективности методов классификации, независящую от конкретного алгоритма самого метода.

В начале напомним задачу распознавания образов в её простейшей постановке.

Пусть имеются N объектов, каждый из которых обладает некоторым набором признаков (X_1, X_2, ...). Известно, что часть из этих объектов принадлежит к классу A, другая часть – к классу B (случай дихотомии). Это так называемые эталонные объекты. В совокупности эти объекты представляют собой точки в признаковом пространстве и образуют «обучающую выборку». Конкретный алгоритм классификации по этой обучающей выборке пытается отнести каждый объект экзаменуемой выборки объёма n к одному из этих двух классов (так называемое «обучение с учителем»).

Таких алгоритмов классификации достаточно много (метод потенциальных функций, стохастической аппроксимации, таксономия, кластерный анализ и т. д.), но все они, естественно, несовершенны в том смысле, что иногда ошибочно относят тот или иной объект экзаменуемой выборки не к тому классу, к которому он на самом деле принадлежит.

Обозначим через Q_{min} долю правильно распознанных объектов ($0 \le Q_{min} \le 1$), гарантированную самим методом, её мы в дальнейшем и будем называть эффективностью метода распознавания. Как уже говорилось, ни один метод теоретических оценок Q_{min} не даёт.

Поэтому в инженерной практике поступают так: применяют акт распознавания к l экзаменуемым выборкам (каждая объёмом n) и в качестве Q_{min} принимают число:

$$Q_{min} = q = \min_{1<i<l}\{Q_i\},$$

где Q_i – доля правильного распознанных объектов в i-й выборке. Такая оценка никак не может быть признана удовлетворительной хотя бы потому, что не гарантирует для (l+1)-й выборки несоблюдения неравенства $Q_{i+1} < q$.

Для получения надёжной и объективной оценки Q_{mjn} поступим так, как это концептуально принято в современной математической статистике. Выдвинем нулевую гипотезу H_0 о том, что отнесение любого объекта экзаменуемой выборки к тому или иному классу происходит независимым и случайным образом, никак не связанным с набором признаков, а опирается лишь на оценочную информацию о доле объектов типа A (или B) в обучающей выборке, и найдём нижнюю оценку эффективности классификации объектов, осуществляемой данной процедурой.

Сравнивая эту оценку с оценкой, полученной «содержательным» методом (альтернатива H_1), решаем вопрос о значимости величины q, т. е. действительно ли данный метод распознавания работает эффективно или же ничем не отличается от простого случайного гадания.

Случайный процесс распознавания будем реализовывать путём подбрасывания специально сконструированной монеты, которая ложится гербом кверху с вероятностью *w*, и решкой – с вероятностью 1-*w*. Если выпал герб, то мы относим объект, например, к классу *A*, если решка – к классу *B*.

Под успехом будем понимать отнесение объекта к тому классу, к которому он на самом деле принадлежит. В противном случае – неудача.

Обозначив через p вероятность успеха, будем иметь:

$$p = \varphi(w, t) = wt + (1 - w)(1 - t) \quad (1)$$

где t – вероятность того, что наугад взятый объект принадлежит классу *A*, (доля объектов типа *A* в выборке). Если мы сравниваем процесс случайного гадания с результатами любого содержательного метода распознавания, то поскольку мы имеем дело с сериями испытаний Бернулли, речь идёт об оценке величины:

$$S = f(p) = \sum_{i=[qn]}^{n} C_n^i p^i (1-p)^{n-i} \quad (2)$$

где [•] – целая часть.

Более того, желательно найти точную нижнюю грань этой величины по *p*:

Inf $f(p) = Q_{min}$ (случайного распознавания)

Из теории вероятностей известно (например, 1), что

$$f(p) = I_p([qn], n - [qn] + 1),$$

где $$I_x(a,b) = \frac{1}{B(a,b)} \int_0^x Z^{a-1} (1-Z)^{b-1} dZ$$

– неполная бета-функция с параметрами *a* и *b*, (*B*(*a*, *b*) – полная бета-функция с теми же параметрами).

С другой стороны [1] $I_x(a, b)$ есть функция бета-распределения со степенями свободы 2a и 2b и как всякая функция распределения, есть неубывающая по *X* функция:

$$P = \left(\frac{X_1^2}{X_1^2 + X_2^2} \le X \right) = I_x(a,b),\ a = \frac{\nu_1}{2},\ b = \frac{\nu_2}{2}$$

где X_1^2 и X_2^2 – независимые случайные величины, имеющие хи-квадрат распределение со степенями свободы ν_1 и ν_2 соответственно.

Следовательно $f(p)$ есть неубывающая по *p* функция и при любом фиксированном *W** (с учётом (1)):

$$\inf_{p\in[0,1]} f(p) = f\left[\inf_{t\in(t_1,t_2)} \varphi\left(w^*,t\right)\right] \tag{3}$$

где t_1 и t_2 – нижняя и верхняя границы доверительного интервала для t. Дело в том, что на практике значение t (доля объектов типа A или B в обучающей выборке) оценивают с помощью доверительного интервала (t_1, t_2), где t_1 и t_2 определяют, соответственно, решением уравнений [2]:

$$I_{t_1}(\mu, r-\mu+1) = 1-P \tag{4}$$

$$I_{t_2}(\mu+1, r-\mu) = P \tag{5}$$

где P – заданный коэффициент доверия ($0.5 \le P \le 1$), r – общее количество независимых испытаний, μ – число успехов, $I_x(a, b)$ – функция B – распределения.

Кроме того учтём, что равенство (3), строго говоря, справедливо при $½ \le w^* \le 1$ (это следует из (1)). Но распространение его на весь диапазон $w^* \in [0,1]$ делается очень просто. В силу симметрии классов A и B, если $0 \le w^* \le ½$, мы относим объект к классу A, если монета ложится решкой, а не гербом (с вероятностью $w_1 = 1 - w$, где уже $½ \le w_1 \le 1$)

Желая достичь наибольшей эффективности модели Бернулли как метода классификации, необходимо так выбрать параметр w, чтобы обеспечивать $\max_w \varphi(w,t)$ при фиксированном t.

Таким образом задача сводится к отысканию

$$\inf_{t\in(t_1,t_2)} \max_{\varphi\in[0,1]} \varphi(w,t) \tag{6}$$

Рассмотрим 3 возможных случая.

1. $t_2 \le ½$

Тогда $\max_w \varphi(w,t) = \varphi(0,t) = 1-t$ и

$$\inf \max_w \varphi(w,t) = 1-t_2 \tag{7}$$

2. $t_1 \ge ½$

Тогда $\max_w \varphi(w,t) = \varphi(1,t) = t\,u$

$$\inf \max_w \varphi(w,t) = t_1 \tag{8}$$

3. $t_1 < ½ < t_2$

При равновероятном попадании t в любую точку интервала (t_1, t_2) максимальное значение w будет в среднем равно:

$$w_{cp_{\max}} = \frac{\frac{1}{2} - t_1}{t_2 - t_1} \cdot 0 + \frac{t_2 - \frac{1}{2}}{t_2 - t_1} \cdot 1 = \frac{1}{2} \cdot \frac{2t_2 - 1}{t_2 - t_1}$$

и

$$\inf_t \max_w \varphi(w,t) = \frac{4t_1(t_2 - 1) + 1}{2(t_2 - t_1)} \qquad (9)$$

Теперь достаточно в выражение для S вместо p подставить правую часть (7), (8) или (9) (в зависимости от конкретного соотношения объектов A и B в обучающей выборке) и мы получим гарантированную оценку (в виде точной нижней грани) эффективности случайного гадания, как метода распознавания образов, которую можно сравнивать с результатами того или иного конкретного «содержательного» метода.

Вот некоторые численные примеры:

Таблица 1

p	0,5		0,7		0,8	
n	10	20	10	20	10	20
$q \cdot 100$ %	80	70	80	70	80	70
inf S	0,055	0,058	0,383	0,608	0,678	0,913

Напомним, что в инженерной практике H_1 отвергают в пользу H_0 уже при inf $S \geq 0{,}05$, поэтому в данном примере результаты распознавания по «содержательному» методу не могут быть признаны убедительными, причём такое утверждение на практике справедливо с надёжностью 0,95–0,99.

Список литературы

1. Справочник по специальным функциям с формулами, графиками и математическими таблицами. Пер. с англ. / Под ред. М. Абрамовича, И. Стиган. М. : Наука, 1979. 830 с.

2. Большев Л. Н., Смирнов Н. В. Таблицы математической статистики. М. : Наука, 1983. 416 с.

List of references

1. Abramovich M., Stigan I. (editors), *Handbook of special functions with formulas, graphs and mathematical tables*. Moscow, Nauka, 1979, 830 p.

2. Bolshev L. N., Smirnov N. V. *Tables of mathematical statistics*. Moscow, Nauka, 1983, 416 p.