ОБОЗРЕНИЕ
ПРИКЛАДНОЙ И ПРОМЫШЛЕННОЙ
Том 27 МАТЕМАТИКИ Выпуск 1
2020

Д. С. Богданов, **В. О. Мироншкин** (Москва, Лаб. ТВП, Национальный исследовательский университет «Высшая школа экономики»). **О трудоемкости проверки качества классификаторов.**

УДК 519.212.2+004.032.2

Резюме: В работе рассматривается подход к оценке качества классификаторов, основанный на подсчете числа по-разному классифицированных «близких» по признакам объектов и сравнении его с заранее заданным пороговым значением, определяемым конкретными целями и условиями задачи. Вычислена средняя трудоемкость процедуры оценки качества классификатора в модели равновероятного случайного отображения.

Ключевые слова: машинное обучение, задача классификации, критерий качества классификатора, равновероятное случайное отображение.

Введение. В настоящее время технологии машинного обучения и искусственного интеллекта являются наиболее перспективными в области обработки и анализа «больших данных». Одной из основных задач, решаемых с помощью методов машинного обучения, является «задача классификации» [4], возникающая, например, при проведении медицинской диагностики [2], геологической разведки [6], оптическом распознавании текстов [3], синтезе химических соединений [7] и т. д. При разработке новых алгоритмов, решающих «задачу классификации», возникает естественная потребность определения критериев их качества [1].

В настоящей работе рассматривается подход к оценке качества подобных алгоритмов, заключающийся в их тестировании на основе некоторого фиксированного подмножества объектов.

Итак, пусть для произвольного множества X (далее — множество объектов) и произвольного конечного множества Y (далее — множество классов) задан некоторый алгоритм $f\colon X \to Y$, классифицирующий любой объект $x \in X$, реализации которого известны только для конечного подмножества $A = \{x_1, \cdots, x_m\} \in X^m$ (A — обучающая выборка [4]), $m \in \mathbb{N}$.

Замечание 1. Одним из способов формализации множества X является его описание через, так называемое, *признаковое пространство* $D_{f_1} \times \cdots, \times D_{f_n}$, где D_{f_i} — множество допустимых значений признака, а отображения $f_i\colon X \to D_{f_i}$ — соответствующие признаки, $i = 1, \ldots, n$ [5].

При этом вектор $\hat{x} = (f_1(x), \ldots, f_n(x))$ называется *признаковым описанием объекта* $x \in X$.

В качестве признакового пространства рассмотрим множество $\{0,1\}^n$, что соответствует $D_{f_i} = \{0,1\}$, $i = 1, \ldots, n$). В этом случае каждый объект обладает набором из $t \leqslant n$ признаков (признаковое описание объекта — бинарный вектор длины n веса t).

Пусть $|Y| = q \in \mathbb{N}$. Не ограничивая общности, в качестве Y будем рассматривать множество $\{0, 1, \ldots, q-1\}$. В этом случае указанный выше алгоритм (далее — классификатор) имеет вид

$$f\colon \{0,1\}^n \to \{0, \ldots, q-1\}. \tag{1}$$

В ряде практических приложений к классификатору α предъявляется следующее требование: «близкие» по признакам объекты должны классифицироваться одинаково, а «далекие» — по-разному.

Замечание 2. Указанному требованию, например, удовлетворяют медицинские классификаторы болезней: «пациенты с похожими симптомами и результатами анализов, как правило, имеют один и тот же диагноз».

Таким образом, один из подходов к оценке качества классификаторов как раз и состоит в оценке числа по-разному классифицированных «близких» по признакам объектов. В частности, классификатор признается качественным, если указанная характеристика не превосходит некоторого порогового значения, определяемого в условиях конкретной решаемой задачи.

В рамках исследования данного подхода оценим среднее число объектов, отличающихся в фиксированном количестве признаков и используемых для тестирования теоретической модели классификатора, построенного на основе равновероятного случайного отображения вида (1).

Рассмотрим вероятностное пространство случайных классификаторов $(\Omega, \mathcal{F}, \mathbf{P})$, где Ω — множество всех функций $f\colon \{0,1\}^n \to \{0,\ldots,q-1\}$, алгебра событий $\mathcal{F}$ – множество всех подмножеств Ω, а вероятностная мера $\mathbf{P}$ задана следующим образом:

$$\mathbf{P}(f) = \frac{1}{q^{2^n}} \quad \forall f \in \Omega. \tag{2}$$

Далее будем использовать следующие определения (в определениях функция f считается детерминированной).

О п р е д е л е н и е 1. Расстоянием Хэмминга χ между двумя произвольными бинарными векторами $\bar{a} = (a_1,\ldots,a_n), \bar{b} = (b_1,\ldots,b_n)$ длины $n \in \mathbb{N}$ называется число координат, в которых эти векторы отличаются:

$$\chi\left(\bar{a},\bar{b}\right) = \sum_{i=1}^{n} I\left\{a_i \neq b_i\right\},$$

где $I\left\{a_i \neq b_i\right\}$ — индикатор события $\left\{a_i \neq b_i\right\}$.

О п р е д е л е н и е 2. Для произвольного $d \in \{1,\ldots,n\}$ назовем (f,d)-*парой* произвольную пару $(\bar{a},\bar{b}) \in \{0,1\}^n \times \{0,1\}^n$, для которой $\chi\left(\bar{a},\bar{b}\right) = d$ и $f(\bar{a}) \neq f\left(\bar{b}\right)$.

Для произвольных фиксированных $(\bar{a},\bar{b}) \in \{0,1\}^n \times \{0,1\}^n$ через $\Psi_f^{(d)}$ обозначим множество всех (f,d)-пар, а через $\xi_f^{(d)}$ — случайную величину, равную мощности множества $\Psi_f^{(d)}$ при случайном выборе f.

Поскольку распределение (2) на Ω индуцирует равновероятное распределение на множестве значений случайной функции f, для любого $d \in \{1,\ldots,n\}$ и любых фиксированных $(\bar{a},\bar{b}) \in \{0,1\}^n \times \{0,1\}^n$, для которой $\chi\left(\bar{a},\bar{b}\right) = d$ и $f(\bar{a}) \neq f\left(\bar{b}\right)$, выполняется равенство

$$\mathbf{P}\left\{(\bar{a},\bar{b}) \text{ является } (f,d)-\text{парой}\right\} = 1 - \frac{1}{q},$$

и поэтому справедлив следующий результат.

Утверждение 1. *Пусть случайная булева функция $f \in \Omega$ имеет распределение (2) на Ω. Тогда для любого $d \in \{1,\ldots,n\}$ справедливо равенство*

$$\mathbf{E}\xi_f^{(d)} = \frac{1}{q} 2^{n-1} C_n^d (q-1). \tag{3}$$

Формула (3) позволяет оценить среднюю трудоемкость процедуры тестирования случайного классификатора на векторах с заданным значением расстояния Хэмминга $d \in \{1,\ldots,n\}$.

Так если процедура тестирования заключается в последовательном опробовании (f,d)-пар, начиная с $d=1$ и заканчивая $d=t\leqslant n$, средняя трудоемкость $\mathbf{E}\xi_f^{(\leqslant t)}$ такой процедуры определяется выражением

$$\mathbf{E}\xi_f^{(\leqslant t)}=\frac{1}{q}2^{n-1}(q-1)\sum_{i=1}^{t}C_n^i. \tag{4}$$

В частности, при $t=n$ величина $\mathbf{E}\xi_f^{(\leqslant n)}$ достигает своего максимума и составляет

$$\mathbf{E}\xi_f^{(\leqslant n)}=C_{2^n}^2\left(1-\frac{1}{q}\right).$$

СПИСОК ЛИТЕРАТУРЫ

1. *Клячкин В. А.*, *Шунина В. Н.*, *Алексеева Ю. С.*. Критерии качества работы классификаторов. — Вестник УлГТУ, 2015, 70(2).
2. *Хурса Р. В*, *Войтикова М. В.*, *Войтович А. П.* Применение интеллектуального анализа данных для классификации гемодинамических состояний. — Артериальная гипертензия, 2015, 43(5), с. 36–42.
3. *Барташевич А. А.*, *Казаков С. Г.* Оптическое распознавание музыкальной транскрипции с помощью методов машинного обучения. — Современные инновации, 2019, 33(5), с. 7–10.
4. *Флах П.* Машинное обучение. ДМК Пресс, 2015.
5. *Воронцов К. В.* Курс лекций «Математические методы обучения по прецедентам (теория обучения машин)», 2019.
6. *Курганов Д. В.* Об одном методе классификации нефтяного месторождения с использованием комплекса геолого-промысловых данных и машинного обучения. — Вестник НГУ, 2020., 18(1), p. 27–35.
7. *Jensen Klavs F.*, *Coley Connor W.*, *Green William H.* Machine learning in computer-aided synthesis planning. Accounts of chemical research, 2018, 51(5), с. 1281–1289,

УДК 519.87+519.24

Bogdanov D. S., Mironkin V. O. (Moscow, TVP Laboratory, National Research University Higher School of Economics)**. On complexity of checking the quality of classifiers.**

Abstract: The approach to estimation the quality of classifiers, based on counting the number of differently classified “close” by features objects and comparing it with a determined by specific goals and conditions of the problem threshold value is considered. The average complexity of the procedure for estimation the quality of the classifier in the model of equiprobable random mapping is obtained.

Keywords: machine learning, classification problem, classifier quality criterion, equiprobable random mapping.