ОБ ОЦЕНКАХ ОЖИДАЕМОГО КАЧЕСТВА ПРИЗНАКОВ

М. М. БОНГАРД, М. Н. ВАЙНЦВАЙГ
(МОСКВА)

Во многих алгоритмах распознавания классификация объектов производится с помощью некоторого числа признаков каждого класса, которые отбираются в процессе обучения на множестве примеров [1—4]. При отборе таких признаков естественно стремиться к тому, чтобы вероятность ошибки при распознавании каждым признаком в отдельности была возможно меньшей.

В связи с этим возникает задача оценки качества отбираемого признака.

1. Рассмотрим множество объектов $\mathfrak{A} = \{a\}$, которое для простоты будем считать конечным. Это ограничение несущественно для дальнейших результатов работы, однако позволяет избежать оговорок, связанных, например, с измеримостью подмножеств множества $\mathfrak{A}$.

Рассмотрим некоторое разбиение ω^ν множества $\mathfrak{A}$ на два непересекающихся подмножества (класса) $\mathfrak{A}_1^\nu$ и $\mathfrak{A}_2^\nu$:

$$\omega^\nu = (\mathfrak{A}_1^\nu, \mathfrak{A}_2^\nu);\ \mathfrak{A}_1^\nu \cap \mathfrak{A}_2^\nu = \varnothing;\ \mathfrak{A}_1^\nu \cup \mathfrak{A}_2^\nu = \mathfrak{A}.$$

(Аналогичным образом могут быть рассмотрены разбиения на любое конечное число классов.)

Будем считать, что с каждым разбиением ω^ν на множестве $\mathfrak{A}$ вводится вероятностная мера λ^ν, определенная для всех объектов $a \in \mathfrak{A}$, а следовательно (в силу конечности $\mathfrak{A}$), и для всех подмножеств $A \subset \mathfrak{A}$. Множество $\mathfrak{A}$ с заданным на нем разбиением ω^ν и мерой λ^ν будем называть *задачей*.

Будем считать, что множество примеров для обучения образуется следующим образом. Выбирая независимо один от другого N_1 объектов класса $\mathfrak{A}_1^\nu$ с мерой $\lambda^\nu(a)/\lambda^\nu(\mathfrak{A}_1^\nu)$, получаем множество A_1 примеров класса $\mathfrak{A}_1^\nu$. Аналогично, выбирая N_2 объектов класса $\mathfrak{A}_2^\nu$ с мерой $\lambda^\nu(a)/\lambda^\nu(\mathfrak{A}_2^\nu)$, получаем множество A_2 примеров класса $\mathfrak{A}_2^\nu$. Для множества примеров $A = A_1 \cup A_2$ считается известной их принадлежность к соответствующим классам.

2. *Признаком* будем называть произвольный предикат, определенный на всех объектах множества $\mathfrak{A}$. Каждому признаку β соответствует множество истинности, которое обозначим через B. Будем говорить, что *объект a обладает признаком* β или что признак β характеризует объект a, если $a \in B$. Будем считать, что задана система признаков $\{\beta^j\}$.

Рассмотрим произвольный признак β^j из этой системы. Обозначим через $\varphi^{\nu j}$ меру множества истинности признака β^j при разбиении ω^ν, а через $p_i^{\nu j}$ — вероятность того, что произвольный объект, обладающий

признаком β^j, принадлежит в данной задаче классу $\mathfrak{A}_i$ $(i = 1, 2)$:

$$\varphi^{\nu j} = \lambda^{\nu}(B^j); \qquad p_i^{\nu j} = \frac{\lambda^{\nu}(B^j \cap \mathfrak{A}_i)}{\lambda^{\nu}(B^j)}.$$

Иногда для краткости индексы i, j, ν будут опускаться.

Во время обучения алгоритм отбирает (пользуясь каким-то критерием) некоторое количество признаков. При классификации часть отобранных признаков будет использоваться в качестве доводов за то, что неизвестный объект принадлежит классу $\mathfrak{A}_1$, а часть — в качестве доводов за то, что он принадлежит классу $\mathfrak{A}_2$. Эти признаки будут называться соответственно *признаками класса* $\mathfrak{A}_1$ *и класса* $\mathfrak{A}_2$. Если β является признаком класса $\mathfrak{A}_i$, то p_i^j будем называть *вероятностью правильной его работы.*

Наша цель, пользуясь информацией, содержащейся в множестве примеров, отобрать для каждого класса $\mathfrak{A}_i$ такие признаки β^j, для которых были бы возможно бóльшими вероятности p_i^j «правильной работы».

Зная числа k_1 и k_2 примеров соответственно первого и второго классов, охарактеризованных признаком β, и воспользовавшись формулой Байеса, можно оценить вероятность $\eta(p_i^*/k_1; k_2)$ того, что признак β, считающийся признаком класса $\mathfrak{A}_i$, имеет вероятность «правильной работы», равную p_i^*,

$$\eta(p_i^*/k_1; k_2) = \frac{\sum\limits_{p_i = p_i^*} f(\varphi, p_1, p_2)\,\zeta(k_1; k_2/\varphi, p_1, p_2)}{\sum\limits_{\varphi, p_1, p_2} f(\varphi, p_1, p_2)\,\zeta(k_1; k_2/\varphi, p_1, p_2)}, \tag{1}$$

где

$$\zeta(k_1; k_2/\varphi, p_1, p_2) =$$

$$= C_{N_1}^{k_1}\left(\frac{\varphi p_1}{\lambda(\mathfrak{A}_1)}\right)^{k_1}\left(1-\frac{\varphi p_1}{\lambda(\mathfrak{A}_1)}\right)^{N_1-k_1} C_{N_2}^{k_2}\left(\frac{\varphi p_2}{\lambda(\mathfrak{A}_2)}\right)^{k_2}\left(1-\frac{\varphi p_2}{\lambda(\mathfrak{A}_2)}\right)^{N_2-k_2} \tag{2}$$

и $f(\varphi, p_1, p_2)$ — априорное распределение вероятностей того, что рассматриваемый признак имеет меру φ и произвольный объект, обладающий этим признаком, с вероятностью p_1 принадлежит классу $\mathfrak{A}_1$ и с вероятностью $p_2 = 1 - p_1$ — классу $\mathfrak{A}_2$.

Качество признака естественно оценивать математическим ожиданием вероятности p_i^*, которое равно

$$M\left(p_i^*/k_1; k_2\right) = \sum_{p_i^*} \eta\left(p_i^*/k_1; k_2\right) p_i^*. \tag{3}$$

Таким образом, для оценки признаков необходимо знать не только параметры k_1 и k_2, которые определяются на множестве примеров, но и априорное распределение $f(\varphi, p_1, p_2)$.

Основная трудность состоит в том, что для большинства практических задач распределение $f(\varphi, p_1, p_2)$ обычно бывает неизвестно. Это обстоятельство естественно приводит к большому произволу в выборе оценки признаков при отборе, что в свою очередь снижает качество классификации посредством отобранных признаков.

Ниже приводятся два различных подхода к решению задачи об оценке признаков, позволяющих в какой-то мере обойти указанную трудность.

3. П е р в ы й п о д х о д состоит в том, что выделяется такой класс признаков и такой (достаточно широкий) класс распределений $f(\varphi, p_1, p_2)$,

что оценка признаков становится монотонной функцией параметров k_1 и k_2 *).

Рассмотрим всевозможные разбиения ω^ν множества $\mathfrak{A}$ на классы $\mathfrak{A}_1^\nu$ и $\mathfrak{A}_2^\nu$ такие, что $\lambda^\nu(\mathfrak{A}_i^\nu) = \lambda(\mathfrak{A}_i) = C_i = \text{const}$ для всех разбиений. Обозначим через $\Omega = \{\omega^\nu\}$ множество всех таких разбиений. Будем считать, что на множестве Ω задано распределение вероятностей $\psi(\omega^\nu)$, определенное для всех разбиений ω^ν, или, что то же самое, для всех задач. Для каждого признака β^j распределение ψ определяет распределение $f^j(\varphi, p_1, p_2)$ вероятностей появления разбиения, на котором признак β^j имеет характеристики φ, p_1, p_2. В дальнейшем будем считать, что распределения $f^j(\varphi, p_1, p_2)$ для всех признаков β^j равны ($f^j(\varphi, p_1, p_2) = f(\varphi, p_1, p_2)$).

Признак β будем называть *достаточным признаком класса* $\mathfrak{A}_i$ на множестве $A \subset \mathfrak{A}$, если им обладают некоторые объекты $a' \in \mathfrak{A}_i \cap A$ и не обладает ни один объект $a'' \in A \setminus (\mathfrak{A}_i \cap A)$. Очевидно, что если множество A совпадает со всем множеством $\mathfrak{A}$, то понятие достаточности на множестве A совпадает с общепринятым понятием достаточности.

В дальнейшем (в рамках первого подхода) при отборе признаков мы будем рассматривать лишь признаки каждого из классов, достаточные на множестве примеров. Пользоваться такими признаками часто целесообразно по нескольким причинам. Во-первых, при этом гарантируется правильность классификации объектов, входящих в множество примеров. Во-вторых, для некоторых типов признаков существуют методы [4] сокращенного отбора достаточных признаков, что дает возможность рассматривать более широкую систему признаков. В-третьих, как будет показано ниже, при сравнительно общих предположениях относительно распределения $f(\varphi, p_1, p_2)$ среди признаков, достаточных на множестве примеров, можно отбирать признаки, имеющие наибольшую оценку. Такой отбор возможен потому, что $M(p/k_1; 0)$ является для таких признаков монотонной функцией k_1.

Преобразуем набор $(\varphi^{\nu j}, p_1^{\nu j}, p_2^{\nu j})$ вероятностных характеристик признака β^j на разбиении ω^ν, обозначив через $\gamma_i^{\nu j} = \lambda^\nu(B^j \cap \mathfrak{A}_i) = \varphi^{\nu j} p_i^{\nu j}$ меру множества истинности признака β^j на классе $\mathfrak{A}_i$. Получим новый набор характеристик $(\gamma_1^{\nu j}, \gamma_2^{\nu j})$. Такое преобразование, очевидно, взаимно однозначно:

$$\varphi^{\nu j} = \sum_i \gamma_i^{\nu j}; \quad p_i^{\nu j} = \frac{\gamma_i^{\nu j}}{\sum\limits_i \gamma_i^{\nu j}},$$

поэтому задание распределения $f(\varphi, p_1, p_2)$ эквивалентно заданию распределения $F(\gamma_1, \gamma_2)$. Справедлива следующая

Т е о р е м а. *Если функции распределения мер* γ_i *для разных классов* $\mathfrak{A}_i$ *независимы, т. е.* $F(\gamma_1, \gamma_2) = F_1(\gamma_1) F_2(\gamma_2)$, *то математическое ожидание* $M(p_i/k)$ *вероятности* p_i *правильной работы признака, оказавшегося на множестве примеров достаточным признаком класса* $\mathfrak{A}_i$, *является монотонно неубывающей функцией числа* k *примеров, обладающих этим признаком.*

З а м е ч а н и е 1. Далеко не для всякого распределения $F(\gamma_1, \gamma_2)$ (а следовательно, и $f(\varphi, p_1, p_2)$) математическое ожидание $M(p_i/k)$ монотонно не убывает с ростом k. Примером может служить случай, когда равновероятны два типа признаков. Признаки первого типа имеют большую меру φ и вероятность p правильной работы, равную 1/2. Признаки второго типа имеют малую меру φ и вероятность правильной работы,

*) Другой класс распределений, при которых оценка признака также становится монотонной функцией параметров k_1 и k_2, приводится в работе [4].

равную 1. Вероятности признаков других типов равны 0. При достаточно малой мере признаков второго типа математическое ожидание $M(p/k)$ будет монотонно убывать с ростом k.

Доказательство теоремы. Для конкретности будем исходить из предположения, что рассматриваемый признак является признаком класса $\mathfrak{A}_1$.

Очевидно, достаточно доказать, что при выполнении условий теоремы разность

$$\Delta M(p/k) = M(p/k+1) - M(p/k)$$

неотрицательна для любого k.

В силу (1), (2), (3)

$$\Delta M(p/k) = \frac{1}{A} \sum_{i,\,j,\,l,\,m} F(\gamma_1^i,\ \gamma_2^j)\, F(\gamma_1^l,\ \gamma_2^m)\, R_{ijlm}\, \frac{\gamma_1^i(\gamma_1^i - \gamma_1^l)}{\gamma_1^i + \gamma_2^j},$$

где

$$R_{ijlm} = \frac{1}{c_1}\left(\frac{\gamma_1^i}{c_1}\right)^k \left(\frac{\gamma_1^l}{c_1}\right)^k \left(1-\frac{\gamma_1^i}{c_1}\right)^{N_1-k-1} \left(1-\frac{\gamma_1^l}{c_1}\right)^{N_1-k-1} \times$$

$$\times \left(1-\frac{\gamma_2^j}{c_2}\right)^{N_2} \left(1-\frac{\gamma_2^m}{c_2}\right)^{N_2};$$

$$A = \left[\sum_{l,\,m} F(\gamma_1^l \gamma_2^m) \left(\frac{\gamma_1^l}{c_1}\right)^{k+1} \left(1-\frac{\gamma_1^l}{c_1}\right)^{N_1-k-1} \left(1-\frac{\gamma_2^m}{c_2}\right)^{N_2}\right] \times$$

$$\times \left[\sum_{l,\,m} F(\gamma_1^l,\ \gamma_2^m) \left(\frac{\gamma_1^l}{c_1}\right)^{k} \left(1-\frac{\gamma_1^l}{c_1}\right)^{N_1-k} \left(1-\frac{\gamma_2^m}{c_2}\right)^{N_2}\right];$$

$$c_1 = \lambda(\mathfrak{A}_1);\ c_2 = \lambda(\mathfrak{A}_2) \quad \text{и} \quad F(\gamma_1^t,\ \gamma_2^q) = \sum_{\substack{\gamma_1^{vs}=\gamma_1^t \\ \gamma_2^{vs}=\gamma_2^q}} F(\gamma_1^{vs},\ \gamma_2^{vs}).$$

Воспользовавшись условием независимости $F(\gamma_1,\ \gamma_2) = F_1(\gamma_1)\, F_2(\gamma_2)$ и объединив в сумме для $\Delta M(p/k)$ все члены с $F_1(\gamma_1^i)\, F_2(\gamma_2^j)\, F_1(\gamma_1^l)\, F_2(\gamma_2^m)$, получим

$$\Delta M(p/k) = \frac{1}{A} \sum_{i,\,j,\,l,\,m} F_1(\gamma_1^i)\, F_2(\gamma_2^j)\, F_1(\gamma_1^l)\, F_2(\gamma_2^m)\, R_{ijlm} Q_{ijlm},$$

где

$$Q_{ijlm} = (\gamma_1^i - \gamma_2^l)\left[\frac{\gamma_1^i}{\gamma_1^i+\gamma_2^j} - \frac{\gamma_1^l}{\gamma_1^l+\gamma_2^j} + \frac{\gamma_1^i}{\gamma_1^i+\gamma_2^m} - \frac{\gamma_1^l}{\gamma_1^l+\gamma_2^m}\right] =$$

$$= (\gamma_1^i - \gamma_1^l)^2 \left[\frac{\gamma_2^j}{(\gamma_1^i+\gamma_2^j)(\gamma_1^l+\gamma_2^j)} + \frac{\gamma_2^m}{(\gamma_1^i+\gamma_2^m)(\gamma_1^l+\gamma_2^m)}\right] \geqslant 0,$$

а так как $R_{ijlm} \geqslant 0$ и $A \geqslant 0$, то и $\Delta M(p/k) \geqslant 0$. Теорема доказана.

Замечание 2. Легко показать, что $\Delta M(p/k) = 0$ лишь в случае, когда $F_1(\gamma_1)$ и $F_2(\gamma_2)$ являются δ-функциями.

4. При отборе признаков нас часто не интересуют точные значения вероятностей правильной работы признаков.

Достаточно лишь упорядочить эти вероятности с тем, чтобы выбирать признаки, имеющие наибольшую вероятность правильной работы.

Для признаков, достаточных на множестве примеров, доказанная теорема рекомендует отбирать признаки с бóльшими k. Напомним, что эта рекомендация справедлива при выполнении следующих условий:

а) Меры класса во всех задачах равны между собой: $\lambda^v(\mathfrak{A}_i) = C_i$.

б) Распределения f по задачам для всех признаков равны между собой: $f^j(\varphi, p_1, p_2) = f(\varphi, p_1, p_2)$.

в) Распределения мер γ для разных классов независимы: $F(\gamma_1, \gamma_2) = F_1(\gamma_1)\, F_2(\gamma_2)$.

В работе [4] была выведена та же рекомендация при несколько других предположениях.

Доказанная там теорема справедлива в случае равенства мер φ для всех признаков ($\varphi^j = \varphi$).

5. Второй подход заключается в том, что в процессе отбора признаков можно приближенно найти распределение $f^\nu(\varphi, p_1, p_2)$ для данной задачи. Для конкретности будем считать, что $\frac{N_1}{N_2} = \frac{\lambda^\nu(\mathfrak{A}_1^\nu)}{\lambda^\nu(\mathfrak{A}_2^\nu)}$. Допустим, мы много раз производим выбор множества примеров для обучения, соблюдая это соотношение. Тогда для некоторого признака β^j $\frac{M(k_1^j + k_2^j)}{N_1 + N_2} = \varphi^j$. Аналогично

$$\frac{M(k_1^j)}{M(k_1^j + k_2^j)} = p_1^j \quad \text{и} \quad \frac{M(k_2^j)}{M(k_1^j + k_2^j)} = p_2^j.$$

Здесь M — математические ожидания соответствующих величин. В большинстве практических случаев нет возможности многократно производить выбор множества примеров для обучения. Однако если N_1 и N_2 не слишком малы, то в качестве некоторого приближения можно пользоваться просто выражениями:

$$\varphi^j \approx \frac{k_1^j + k_2^j}{N_1 + N_2}, \qquad p_1^j \approx \frac{k_1^j}{k_1^j + k_2^j}, \qquad p_2^j \approx \frac{k_2^j}{k_1^j + k_2^j}. \tag{4}$$

Пусть алгоритм, производя отбор признаков, вычисляет, кроме того, для каждого проверяемого признака φ, p_1 и p_2 по формулам (4). К концу обучения появляется возможность приближенно построить распределение $f(\varphi, p_1, p_2)$ для всех доступных алгоритму признаков.

Если отбор признаков производится с помощью некоторого критерия, использующего только числа k_1 и k_2, то для каждого признака β^j есть вероятность $L_i^j = L_i(\varphi^j, p_1^j, p_2^j)$ того, что он будет отобран в качестве признака класса $\mathfrak{A}_i$. Функция $L_i(\varphi, p_1, p_2)$ зависит только от критерия отбора (не зависит от задачи) *) и благодаря этому может быть вычислена заранее.

Таким образом, к концу процесса обучения появляется возможность оценить ожидаемое качество отобранных признаков:

$$M_i(p_i) = \frac{\sum\limits_{\varphi, p_1, p_2} p_i \varphi f(\varphi, p_1, p_2)\, L_i(\varphi, p_1, p_2)}{\sum\limits_{\varphi, p_1, p_2} \varphi f(\varphi, p_1, p_2)\, L_i(\varphi, p_1, p_2)}, \tag{5}$$

*) Например, если в качестве признаков первого класса отбираются признаки, характеризующие не менее nk_2 объектов из A_1 (k_2 — число охарактеризованных объектов из A_2), то

$$L_1(\varphi, p_1, p_2) = \sum_{k_1/k_2 \geqslant n} C_{N_1}^{k_1} X^{k_1} (1-X)^{N_1 - k_1} C_{N_2}^{k_2} Y^{k_2} (1-Y)^{N_2 - k_2},$$

где

$$X = \frac{(N_1 + N_2)\varphi p_1}{N_1}, \qquad Y = \frac{(N_1 + N_2)\varphi p_2}{N_2}.$$

где через $M_i(p_i)$ обозначено математическое ожидание вероятности правильной работы признаков класса $\mathfrak{A}_i$.

Пользуясь формулами (4), мы предполагали, что

$$\frac{N_1}{N_2} = \frac{\lambda^{\nu}(\mathfrak{A}_1^{\nu})}{\lambda^{\nu}(\mathfrak{A}_2^{\nu})}.$$

Однако для возможности произвести оценку $M_i(p_i)$ вовсе не обязательно, чтобы числа примеров каждого класса данных для обучения относились между собой так же, как будут относиться при узнавании числа объектов классов $\mathfrak{A}_1$ и $\mathfrak{A}_2$. Можно сделать произвольное предположение об отношении $\frac{\lambda^{\nu}(\mathfrak{A}_1^{\nu})}{\lambda^{\nu}(\mathfrak{A}_2^{\nu})}$ и произвести оценку для этого ожидаемого при узнавании отношения. При этом очевидным образом изменятся формулы (4) и функции $L_i(\varphi, p_1, p_2)$.

6. До сих пор, говоря о втором подходе, мы считали, что критерий отбора признаков заранее задан (функции $L_i(\varphi, p_1, p_2)$ зафиксированы). В этом случае мог возникнуть вопрос лишь об оценке ожидаемого качества признаков (ожидаемого в данной конкретной задаче, характеризуемой распределением $f(\varphi, p_1, p_2)$). Однако знание распределения $f(\varphi, p_1, p_2)$ открывает и другую возможность. Пусть алгоритм обладает несколькими различными критериями отбора признаков. Тогда, найдя $f(\varphi, p_1, p_2)$ и зная функции L для всех своих критериев, алгоритм может на первом этапе работы с помощью (5) выбрать критерий, хороший для данной задачи, и лишь после этого, пользуясь найденным критерием, начать отбирать признаки.

Вообще говоря, критерий отбора может использовать не только числа k_1 и k_2. Например, в большинстве задач медицинской и технической диагностики приходится, кроме k_1 и k_2, учитывать «степень новизны» (независимости) проверяемых признаков. В таких случаях выбор критерия осложняется двумя обстоятельствами. Во-первых, выбор нужно осуществлять среди весьма большого количества критериев (комбинаций разных вариантов проверки «индивидуальных качеств» признаков и вариантов проверки «новизны» признаков). Во-вторых, функции L перестают быть независимыми от задачи и, кроме того, изменяются в ходе отбора признаков. Из-за этого расчет становится намного сложнее.

На практике обычно используется компромиссный прием. Вместо проверки всех возможных критериев подбирается отдельно «фильтр индивидуальных качеств» признаков (проверяющий только k_1 и k_2) и отдельно «фильтр новизны признаков». Знание распределения $f(\varphi, p_1, p_2)$ дает возможность выбрать первый из этих фильтров.

Весьма неприятная возможность неравномерной охарактеризованности различных объектов остается в силе при любом фильтре по k_1 и k_2. Для борьбы с ней необходимы какие-нибудь виды «фильтров новизны» (проверка новизны аргументов признака, проверка новизны множества $(A_1 \cup A_2) \cap B^j$, процесс «доучивания» и т. п.), см. [4, 5].

* *
*

В настоящее время существует много узнающих программ, в которых автоматизирован процесс отбора признаков. Критерий отбора назначается человеком-оператором часто в результате большой предварительной работы. Оценки ожидаемого качества признаков открывают возможность следующего этапа — автоматизации построения фильтра признаков, «подогнанного» под данную конкретную задачу.

ЛИТЕРАТУРА

1. Бонгард М. М., Моделирование процесса узнавания на цифровой счетной машине, Биофизика **2**, 1961.
2. Браиловский В. Л., Об одном методе распознавания объектов, описываемых несколькими параметрами, и о возможности его применений, Автоматика и телемеханика, № 12, 1962.
3. Браиловский В. Л., Лунц А. Л., Формулировка задачи распознавания объектов со многими параметрами и методы ее решения, Изв. АН СССР, Техническая кибернетика, № 1, 1964.
4. Вайнцвайг М. Н., Об одном алгоритме распознавания двоичных кодов, Проблемы передачи информации, № 3, 1966.
5. Бонгард М. М., Вайнцвайг М. Н., Губерман Ш. А., Извекова М. Л., Смирнов М. С., Использование обучающейся программы для выявления нефтеносных пластов, Геология и геофизика, № 6, 1966.

Поступило в редакцию 19 I 1967