MTA SZTAKI Közlemények 35/1986 p. 135-146

ОЦЕНКА ЭФФЕКТИВНОСТИ КОМБИНАТОРНОГО АЛГОРИТМА РАСПОЗНАВАНИЯ ОБРАЗОВ

Нго Куок Тао, Хоанг Кием

Институт информатики и кибернетики
Ханой - Вьетнам

В этой статье будет показана оценка ошибочной вероятности комбинаторного алгоритма распознавания образов для N исходных алгоритмов и утверждено, что если каждый исходный алгоритм имеет ошибочную вероятность $0 < \varepsilon < 1/2$, то

$$R_N = \sum_{k>\lfloor N/2 \rfloor} \binom{N}{k} \varepsilon^k (1-\varepsilon)^{N-k} \leq \exp(-N.c(\varepsilon)) \qquad /1/$$

где $\lfloor x \rfloor$ - наибольшее целое число, не превосходящее x,

$$c(\varepsilon) = \ln\left(\frac{1}{2\sqrt{\varepsilon(1-\varepsilon)}}\right) > 0 \ .$$

Более того, $\forall \varepsilon \quad 0 < \varepsilon < 1/2 \quad R_N < \varepsilon$ /2/

Из этого следует, что комбинаторный алгоритм лучше каждого данного алгоритма.

Когда исходные алгоритмы A_i имеют ошибочные вероятности ε_i соответственно, то ошибочная верояность комбинаторного алгоритма удовлетворяет неравенству

$$R_N \le \frac{1}{N} \cdot \left(\frac{1}{1 - 2 \max\limits_{1\le i\le N} \varepsilon_i}\right)^2 \quad \text{причем,} \qquad /3/$$

$$R_N \le \frac{4}{N^3} \cdot \left(\frac{\sum\limits_{i=1}^{N} \sqrt{\varepsilon_i(1-\varepsilon_i)}}{1-2\sum\limits_{i=1}^{N} \varepsilon_i/N}\right)^2 . \qquad /3.1/$$

1. ВВЕДЕНИЕ

Комбинаторный метод алгоритма распознавания образов используется для получения нового алгоритма распознавания с эффективностью большей чем эффективность каждого исходного алгоритма, или иначе говоря, получаемый алгоритм имеет вероятность меньше чем вероятность каждого данного алгоритма.

Этот метод основывается на правиле большинственных решений. Он формулируется следующим образом [1]:

Пусть даны N алгоритмов распознавания образов $A_1,\dots,A_N$, классифицирующие объекты $\{s_i \mid i = \overline{1, n}\}$ на два класса K_1, K_2. Определим матрицу $\| a^k_{ij} \|_{n\times 2}$, $k = \overline{1,N}$ как следующее:

$$a^k_{ij} = \begin{cases} 1 & \text{если алгоритм } A_k \text{ относит объект } s_i \text{ к классу } K_j \\ 0 & \text{в обратном случае} \end{cases}$$

Тогда, комбинаторный алгоритм распознания образов, основывающийся на правиле большинственных решений, определит

$$D_j(s_i) = \begin{cases} 1 & \text{если } \sum\limits_{k=1}^{N} a^k_{ij} > \lfloor N/2 \rfloor \\ 0 & \text{в обратном случае} \end{cases}$$

где $D_j(s_i) = 1$ значит, что комбинаторный алгоритм распознавания образов относит s_i к классу K_j, $i = \overline{1,n}$, $j = \overline{1,2}$.

Когда алгоритм $A_1, \ldots, A_N$ независим и каждый из них имеет ошибочную вероятность $\varepsilon > 0$, то ошибочная вероятность комбинаторного алгоритма определяется формулой:

$$R_N = \sum_{k > \lfloor N/2 \rfloor} \binom{N}{k} \varepsilon^k (1-\varepsilon)^{N-k} \qquad /4/$$

В параграфе § 2 этой статьи будет доказано, что $\forall \varepsilon: 0 < \varepsilon \leq 1/2, R_N < \varepsilon$ /это неравенство показывает, что плучаемый комбинаторный алгоритм лучше каждого исходного алгоритма/ и

$$R_N \leq \exp\left(-N.\ln\left(1/2\sqrt{\varepsilon(1-\varepsilon)}\right)\right)$$

/Благодаря этому неравенству будет получен следующие интересный результат:

Если каждый из N данных алгоритмов имеет ошибочную вероятность $0 < \varepsilon < 1/2$, что скорость схождения R_N к нулю имеет экспонентный порядок при $N \to \infty$./

В том случае, когда алгоритмы A_i независимы и имеют ошибочные вероятности ε_i соответственно, то ошибочная вероятность R_N может определяться следующим образом:

$$R_N = \sum_{(\lambda_1, \ldots, \lambda_N) \in TA} \prod_{k=1}^{N} \varepsilon_k^{\lambda_k} (1-\varepsilon_k)^{(1-\lambda_k)},$$

$$TA = \{(\lambda_1, \ldots, \lambda_N) \mid \forall k = \overline{1,N} \quad \lambda_k \in \{0,1\},$$

$$\sum_{k=1}^{N} \frac{\lambda_k}{\sqrt{\varepsilon_k(1-\varepsilon_k)}} > \frac{1}{2} \sum_{k=1}^{N} \frac{1}{\sqrt{\varepsilon_k(1-\varepsilon_k)}}\}$$

Более того, $R_N \le \frac{1}{N} \cdot \left(\frac{1}{1 - 2\max\limits_{1\le i\le N} \varepsilon_i}\right)^2$ и

$$R_N \le \frac{4}{N^3}\left(\frac{\sum\limits_{i=1}^{N} \sqrt{\varepsilon_i(1-\varepsilon_i)}}{1 - 2\sum\limits_{i=1}^{N} \varepsilon_i/N}\right)^2 .$$

Это утверждение будет доказано в параграфе § 3.

2. ОЦЕНКА ОШИБОЧНОЙ ВЕРОЯТНОСТИ КОМБИНАТОРНОГО АЛГОРИТМА ДЛЯ НЕЗАВИСИМЫХ АЛГОРИТМОВ С РАВНЫМИ ОШИБОЧНЫМИ ВЕРОЯТНОСТЯМИ

Теорема 1.

$$R_N = \sum_{k>\lfloor N/2\rfloor} \binom{N}{k} \varepsilon^k (1-\varepsilon)^{N-k}$$ и $R_N < \varepsilon$ для любого ε $0 < \varepsilon \leqslant \frac{1}{2}$.

Доказательство.

Прежде всего, мы видим, что функция $f(\varepsilon) = \varepsilon^a(1-\varepsilon)^b$, $a\ge b>0$, $0<\varepsilon\le 1/2$ строгомонотонная, т.е. $\forall\varepsilon_1\ \forall\varepsilon_2\ \ 0<\varepsilon_1<\varepsilon_2\le\frac{1}{2}$

$f(\varepsilon_1) < f(\varepsilon_2)$. /5/

Поэтому требуемое неравенство эквивалентно

$$T = \sum_{k>\lfloor N/2\rfloor} \binom{N}{k} \varepsilon^{k-1}(1-\varepsilon)^{N-k} < 1 \qquad /6/$$

Из-за того, что $k > \lfloor N/2\rfloor$, то $k - 1 \ge N-k$, в результате неравенства /5/ мы получим

$$\varepsilon^{k-1}(1-\varepsilon)^{N-k} < \left(\tfrac{1}{2}\right)^{k-1} \left(1-\tfrac{1}{2}\right)^{N-k} = \frac{1}{2^{N-1}} .$$

Поэтому

$$T < \sum_{k>\lfloor N/2\rfloor} \binom{N}{k} \cdot \frac{1}{2^{N-1}} = \frac{1}{2^{N-1}} \cdot \sum_{k>\lfloor N/2\rfloor} \binom{N}{k} .$$

При $N = 2m$

$$\sum_{k>\lfloor N/2\rfloor} \binom{N}{k} = \sum_{k=m+1}^{2m} \binom{2m}{k} = \frac{1}{2}\Big(\sum_{k=0}^{m-1} \binom{2m}{k} + \sum_{k=m+1}^{2m} \binom{2m}{k}\Big).$$

Из этого следует, что $T < \frac{1}{2^{2m}}\left(2^{2m} - \binom{2m}{m}\right) = 1 - \frac{\binom{2m}{m}}{2^{2m}} < 1.$

При $N = 2m+1$

$$\sum_{k>\lfloor N/2\rfloor} \binom{2m+1}{k} = \sum_{k=m+1}^{2m+1} \binom{2m+1}{k} = \frac{1}{2} \sum_{k=0}^{2m+1} \binom{2m+1}{k} = 2^{2m}\ .$$

Отсюда следует, что $T < \frac{1}{2^{2m}} \cdot 2^{2m} = 1.$

При $\varepsilon > 0$, последовательность $\{R_N\}$ обладает свойством:

Теорема 2.

$$\forall\varepsilon: 0 < \varepsilon < {}^1/_2,\ R_N = \sum_{k>\lfloor N/2\rfloor} \binom{N}{k} \varepsilon^k (1-\varepsilon)^{N-k} < \exp(-N.c(\varepsilon)),$$

где $c(\varepsilon) = \ln\left(\frac{1}{2\sqrt{\varepsilon(1-\varepsilon)}}\right) > 0$.

Доказательство.

Прежде всего, отмечаем, что при $0 < \varepsilon < {}^1/_2$

$\ln((1-\varepsilon)/\varepsilon) > 0$ и $\ln(1/(2.\sqrt{\varepsilon(1-\varepsilon)})) > 0$.

При $\lambda > 0$ $\exp(\lambda.k) > \exp(\lambda.N/2)$ при $k > \lfloor N/2\rfloor$.

Отсюда следует, что $\exp(-\lambda.N/2) . \exp(\lambda.k) > 1$ /7/

Из-за того, что

$R_N = \sum_{k>\lfloor N/2\rfloor} \binom{N}{k} \varepsilon^k (1-\varepsilon)^{N-k}$ и /7/ получим

$$R_N < \exp(-\lambda.N/2) \sum_{k>\lfloor N/2 \rfloor} \binom{N}{k} \exp(\lambda.k) \cdot \varepsilon^k (1-\varepsilon)^{N-k}$$

$$< \exp(-\lambda.N/2) \cdot \sum_{k=0}^{N} \binom{N}{k} \exp(\lambda.k) \cdot \varepsilon^k \cdot (1-\varepsilon)^{N-k}$$

$$= \exp(-\lambda.N/2)\ (1-\varepsilon+\exp(\lambda).\varepsilon)^N = \left(\frac{1-\varepsilon}{\exp(\lambda/2)} + \varepsilon.\exp(\lambda/2)\right)^N .$$

Положив $\lambda = \ln\left(\frac{1-\varepsilon}{\varepsilon}\right) > 0$, мы получим $\frac{1-\varepsilon}{\exp(\lambda/2)} = \varepsilon.\exp(\lambda/2) = \sqrt{\varepsilon(1-\varepsilon)}$. Поэтому

$$R_N < (2\sqrt{\varepsilon(1-\varepsilon)})^N .$$

Положив $c(\varepsilon) = \ln\left(\frac{1}{2\sqrt{\varepsilon(1-\varepsilon)}}\right)$, имеем $R_N < \exp(-N.c(\varepsilon))$.

/Из факта, что $c(\varepsilon)$ - постоянная следует, что $R_N \to 0$ при $N \to \infty$, R_N имеет скорость схождения экспонентного порядка./

3. ОЦЕНКА ОШИБОЧНОЙ ВЕРОЯНОСТИ КОМБИНАТОРНОГО АЛГОРИТМА ПРИ РАЗЛИЧНЫХ ОШИБОЧНЫХ ВЕРОЯТНОСТЯХ АЛГОРИТМОВ

Правило решений комбинаторного алгоритма определяется следующим образом:

$$D_j(s_i) = \begin{cases} 1 \text{ если } \displaystyle\sum_{k=1}^{N} \frac{a_{ij}^k}{\sqrt{\varepsilon_k(1-\varepsilon_k)}} > \frac{1}{2} \sum_{k=1}^{N} \frac{1}{\sqrt{\varepsilon_k(1-\varepsilon_k)}} \\ 0 \text{ в обратном случае} \end{cases}$$

<u>Теорема 3</u>.

Если исходные алгоритмы A_i независимы и имеют ошибочные вероятности ε_i, соответственно, то ошибочная вероятность R_N может определяться следующим образом:

$$R_N = \sum_{(\lambda_1,\ldots,\lambda_N)\in TA} \prod_{k=1}^{N} \varepsilon_k^{\lambda_k}(1-\varepsilon_k)^{(1-\lambda_k)} ,$$

где $TA = \{(\lambda_1,\ldots,\lambda_N) / \forall k=\overline{1,N} \ \ \lambda_k \in \{0,1\},$

$$\sum_{k=1}^{N} \frac{\lambda_k}{\sqrt{\varepsilon_k(1-\varepsilon_k)}} > \frac{1}{2} \sum_{k=1}^{N} \frac{1}{\sqrt{\varepsilon_k(1-\varepsilon_k)}} \}.$$

Теорема 4.

При условии как в теореме 3 мы получим

$$R_N \leq \frac{1}{N} \cdot \frac{1}{(1-2\max\limits_{1\leq i\leq N} \varepsilon_i)^2} .$$

Доказательство.

Из теоремы 3 следует, что

$$R_N \leq \left(\frac{1}{2}\cdot \sum_{k=1}^{N} \frac{(1-2\varepsilon_k)}{\sqrt{\varepsilon_k(1-\varepsilon_k)}}\right)^{-2} \sum_{(\lambda_1,\ldots,\lambda_N)\in TA} \prod_{k=1}^{N} \varepsilon_k^{\lambda_k}\cdot(1-\varepsilon_k)^{(1-\lambda_k)}$$

$$\left(\sum_{k=1}^{N} \frac{(\lambda_k-\varepsilon_k)}{\sqrt{\varepsilon_k(1-\varepsilon_k)}}\right)^2 .$$

Нетрудно видеть, что

$$\sum_{(\lambda_1,\ldots,\lambda_N)\in TA} \prod_{k=1}^{N} \varepsilon_k^{\lambda_k}(1-\varepsilon_k)^{(1-\lambda_k)} \left(\sum_{k=1}^{N} \frac{(\lambda_k-\varepsilon_k)}{\sqrt{\varepsilon_k(1-\varepsilon_k)}}\right)^2 \leq N.$$

Поэтому

$$R_N \leq \left(\frac{1}{2} \sum_{k=1}^{N} \frac{(1-2\varepsilon_k)}{\sqrt{\varepsilon_k(1-\varepsilon_k)}}\right)^{-2} \cdot N \quad /8/ \quad \leq \left(\frac{1}{2} \sum_{k=1}^{N} 2(1-2\varepsilon_k)\right)^{-2} \cdot N$$

$$\leq [N(1-2\max\limits_{1\leq i\leq N} \varepsilon_i)]^{-2} \cdot N = \frac{1}{N} \cdot \frac{1}{(1-2\max\limits_{1\leq i\leq N} \varepsilon_i)^2} .$$

Теорема 5.

$$R_N \le \frac{4}{N^3} \cdot \frac{\left(\sum_{k=1}^{N} \sqrt{\varepsilon_k(1-\varepsilon_k)}\right)^2}{\left(1 - 2\cdot\frac{1}{N}\cdot\sum_{k=1}^{N} \varepsilon_k\right)^2}$$

Доказательство.

Прежде всего мы докажем следующий результат:

Лемма /Неравенство Чебышева/:

Если последовательности $\{a_i\}$, $\{b_i\}$ удовлетворяют следущим условиям:

$\forall i > j \quad a_i \le a_j$, $b_i \ge b_j$, то

$$\left(\frac{1}{N}\cdot\sum_{i=1}^{N} a_i\right)\left(\frac{1}{N}\cdot\sum_{i=1}^{N} b_i\right) \ge \frac{1}{N}\cdot\sum_{i=1}^{N} a_i b_i .$$

Доказательство

Нетрудно видеть, что $\forall\, i, j \ (a_i - a_j)(b_i - b_j) \le 0$. Поэтому

$$\sum_{i,j} (a_i - a_j)(b_i - b_j) \le 0 \quad \text{или}$$

$$\sum_{i,j} (a_i b_i - a_j b_i - a_i b_j + a_j b_j) \le 0,$$

$$2N \cdot \sum_{i} a_i b_i - 2 \sum_{i,j} a_i b_j \le 0, \text{ или}$$

$$N \sum_{i=1}^{N} a_i b_i \le \left(\sum_{i=1}^{N} a_i\right)\left(\sum_{j=1}^{N} b_j\right).$$

Отсюда мы можем получить неравенство

$$\left(\frac{1}{N}\sum_{i=1}^{N} a_i\right)\left(\frac{1}{N}\sum_{i=1}^{N} b_i\right) \ge \frac{1}{N}\sum_{j=1}^{N} a_i b_i .$$

Сейчас мы докажем теорему:

Из неравенства /8/ следует, что

$$R_N \le \left(\frac{1}{2}\sum_{k=1}^{N}\frac{(1-2\cdot\varepsilon_k)}{\sqrt{\varepsilon_k(1-\varepsilon_k)}}\right)^{-2}\cdot N \ .$$

Используя лемму может получиться неравенство

$$\left(\sum_{k=1}^{N}\frac{(1-2\cdot\varepsilon_k)}{\sqrt{\varepsilon_k(1-\varepsilon_k)}}\right)\left(\sum_{k=1}^{N}\sqrt{\varepsilon_k(1-\varepsilon_k)}\right) \ge N\cdot\left(\sum_{k=1}^{N}(1-2\varepsilon_k)\right)=$$

$$= N^2\left(1-\frac{2}{N}\sum_{k=1}^{N}\varepsilon_k\right) \ .$$

$$R_N \le \frac{4N\left(\sum_{k=1}^{N}\sqrt{\varepsilon_k(1-\varepsilon_k)}\right)^2}{N^4\left(1-\frac{2}{N}\sum_{k=1}^{N}\varepsilon_k\right)^2} = \frac{4}{N^3}\cdot\frac{\left(\sum_{k=1}^{N}\sqrt{\varepsilon_k(1-\varepsilon_k)}\right)^2}{\left(1-2\sum_{k=1}^{N}\varepsilon_k/N\right)^2} \ .$$

Выражаем большую благодарность профессору Бак Хынг Кхангу за постоянную помощь во время работы.

Мы также благодарны нашей коллеге Нгуен Тхать Тзу - из Ханойского политехнического института за хорошие советы и всем нашим коллегам в лаборатории за их симпатию и поддержку нашей работы.

ЛИТЕРАТУРА

[1] Hoang Kiém: Some methods improving the efficiency of pattern recognition algorithms. Computer and artificial intelligence, 3/1984/, N^{o} 4, pp. 347-359.

[2] Gaillat G.: Borne superieure de la probabilité d'eurreur avec la régla de decision majoritaire. Congres AFCET-IRIA, 1978, pp. 230-236.

Alakfelismerési kombinatorikus algoritmusok hatékonyságának becslése

Ngo Kuoc Tao, Hoang Kiem

Összefoglaló

Legyenek $A_1, A_2, \ldots, A_N$ alak-felismerési algoritmusok, amelyek n objektumot két kategóriába sorolnak be, és tegyük fel, hogy /egymástól függetlenül/ ε_i valószinüséggel hibáznak az A_i-k. Az A_i-k segitségével a szerzők egy u.n. "alakfelismerési kombinatorikus algoritmust" definiálnak és annak R_N "hibázási" valószinüségét. A szerzők bebizonyitják, hogy ha $\varepsilon_1 = \varepsilon_2 = \ldots = \varepsilon_N =$ és $0 < \varepsilon \leq \frac{1}{2}$, akkor $R_N < \varepsilon$ és $R_N \leq \exp(-N \ln(1/(2\sqrt{\varepsilon(1-\varepsilon)})))$, az általános esetben pedig $R_N \leq N^{-1}(1-2 \max\limits_{1 \leq i \leq N} \varepsilon_i)^{-2}$ és $R_N \leq 4N^{-3} (\Sigma\sqrt{\varepsilon_i(1-\varepsilon_i)})^2 (1-2\Sigma\, \varepsilon_i/N)^{-2}$.

Estimation for efficiency of pattern-recognition combinatorial algorithms

Ngoc Kuoc Tao, Hoang Kiem

Summary

Let $A_1, A_2, \ldots, A_N$ be pattern-recognition algorithms which classify n objects into two categories. Assume that, independently of each other, the probability that A_i classifies wrongly is ε_i. Using A_i, the authors define a "pattern recognition combinatorial algorithm" and its probability R_N of wrong classification. They prove that if $\varepsilon_1 = \varepsilon_2 = \ldots = \varepsilon_N = \varepsilon$ and $0 < \varepsilon \leq 1/2$ then $R_N < \varepsilon$ and $R_N \leq \exp(-N \ln(1/(2\sqrt{\varepsilon(1-\varepsilon)})))$, and in the general situation $R_N \leq N^{-1}(1-2 \max\limits_{1 \leq i \leq N} \varepsilon_i)^{-2}$ and $R_N \leq 4N^{-3}(\Sigma \sqrt{\varepsilon_i(1-\varepsilon_i)})^2(1 - 2 \Sigma \varepsilon_i/N)^{-2}$.