Tạp chí KHOA HỌC TÍNH TOÁN VÀ ĐIỀU KHIỂN TẬP II (1986), Số 3 (9 - 13)

MỘT SỐ KẾT QUẢ VỀ HIỆU QUẢ CỦA THUẬT TOÁN NHẬN DẠNG TỔ HỢP

NGÔ QUỐC TẠO, HOÀNG KIẾM

1 — ĐẶT VẤN ĐỀ

Phương pháp tổ hợp các thuật toán nhằm nhận được thuật toán có hiệu quả hơn hiệu quả của mỗi thuật toán ban đầu, cụ thể là, nếu xác suất sai lầm của mỗi thuật toán là ε_i thì xác suất sai lầm của thuật toán tổ hợp có thể nhỏ hơn ε_i. Tính cho đến nay đã xuất hiện nhiều công trình đề cập đến thuật toán nhận dạng tổ hợp với các công thức tính R_n [1], [2]. Song đó là những công thức đánh giá chưa được chứng minh chặt chẽ. Trong bài này trình bày một số đánh giá cho R_n chặt hơn một số kết quả đã công bố.

Phương pháp tổ hợp các thuật toán được dựa trên quy tắc quyết định theo đa số được phát biểu như sau:

Giả sử cho trước N thuật toán nhận dạng $\mathcal{A}_1,..., \mathcal{A}_N$ phân loại các dạng $\{S_i | i=1,...,n$ thành 2 lớp K_1 và K_2. Người ta xác định các ma trận $\|a_{ij}^k\|$ $n\times 2$, $k=1,..., N$ như sau:

$$a_{ij}^k = \begin{cases} 1, \text{ nếu thuật toán } \mathcal{A}_k \text{ xếp } S_i \text{ vào lớp } K_j \\ 0, \text{ ngược lại.} \end{cases}$$

Quy tắc quyết định của thuật toán nhận dạng tổ hợp như sau:

$$D_j(S_i) = \begin{cases} 1, \text{ nếu } \sum_{i=1}^{N} a_{ij}^k > E \\ 0, \text{ ngược lại.} \end{cases}$$

Ở đây ký hiệu $E = [LN/2]$ (số nguyên tố lớn nhất không vượt quá $N/2$).

$D_j(S_i) = 1$ có nghĩa là thuật toán nhận dạng tổ hợp xếp S_i vào lớp K_j, $i=1,..., n$; $j = 1,2$. Khi các thuật toán $\mathcal{A}_1 ..., \mathcal{A}_n$ độc lập và mỗi thuật toán có xác suất sai lầm $0<\varepsilon<1/2$ thì xác suất sai lầm của thuật toán nhận dạng tổ hợp được xác định theo công thức sau:

$$R_N = \sum_{k>E} C_N^k \varepsilon^k (1-\varepsilon)^{N-k}$$

Trong II sẽ chỉ ra rằng $\forall \varepsilon$, $0<\varepsilon<1/2$ thì $R_N < \varepsilon$. (Bất đẳng thức này chứng tỏ rằng thuật toán tổ hợp có hiệu quả hơn mỗi thuật toán đã cho). Và $R_N \leqslant \exp\{-N\times C(\varepsilon)\}$ (tốc độ tiến tới 0 của R_N có cấp hàm mũ theo N).

Ta ký hiệu $\vec{\lambda}$ là vectơ có N tọa độ nhận giá trị 0,1 tức là

$$\vec{\lambda} = (\lambda_1, \lambda_2,..., \lambda_N)$$
$$\lambda_k \in \{0.1\},\ k = 1,..., N$$

và

$$TA = \left\{ \vec{\lambda} \Big| \sum_{k=1}^{N} \frac{\lambda_k}{\sqrt{\varepsilon_k(1-\varepsilon_k)}} \geq \frac{1}{2} \sum_{k=1}^{N} \frac{1}{\sqrt{\varepsilon_k(1-\varepsilon_k)}} \right\}$$

Trong trường hợp các thuật toán $\mathcal{A}_i$ độc lập và có xác suất sai lầm ε_i tương ứng, thì xác suất sai lầm R_N có thể xác định theo công thức:

$$R_N = \sum_{\vec{\lambda} \in TA} \prod_{k=1}^{N} \varepsilon_k^{\lambda_k} (1 - \varepsilon_k)^{1-\lambda_k}$$

Hơn nữa

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

Các khẳng định này được chỉ ra trong III.

II – ĐÁNH GIÁ XÁC SUẤT SAI LẦM CỦA THUẬT TOÁN NHẬN DẠNG TỔ HỢP ĐỐI VỚI N THUẬT TOÁN ĐỘC LẬP CÓ CÙNG XÁC SUẤT SAI LẦM GIỐNG NHAU ε.

Bổ đề 1 [1]: Nếu cho trước N thuật toán nhận dạng độc lập có cùng xác suất sai lầm ε thì xác suất sai lầm của thuật toán nhận dạng tổ hợp thỏa mãn:

$$R_N = \sum_{k > E} C_N^k \varepsilon^k (1 - \varepsilon)^{N-k}.$$

Mệnh đề 1: Với ε thỏa mãn $0 < \varepsilon < 1/2$ thì xác suất sai lầm của thuật toán nhận dạng tổ hợp thỏa mãn:

$$R_N < \varepsilon.$$

Chứng minh: Trước hết ta nhận thấy hàm $f_{a,b}(\varepsilon) = \varepsilon^a (1 - \varepsilon)^b$, với $0 < b \leqslant a$, tăng chặt trên $[0, 1/2]$, tức là

$$\forall \varepsilon_1, \varepsilon_2 \text{ thỏa mãn } 0 \leqslant \varepsilon_1 < \varepsilon_2 \leqslant 1/2$$

thì $f_{a,b}(\varepsilon_1) < f_{a,b}(\varepsilon_2)$.

Đặt $T_N = \sum_{k > E} C_N^k \varepsilon^{k-1} (1 - \varepsilon)^{N-k}$, do bổ đề 1 ta có $R_N = T_N . \varepsilon$.

Do nhận xét trên, với $k > E$ hàm $f_{k-1,\, N-k}(\varepsilon)$ tăng chặt trên $[0, 1/2]$. Do đó $f_{k-1,\, N-k}(\varepsilon) \leqslant f_{k-1,\, N-k}(1/2)$.

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

Vì thế

$$T_N < \sum_{k > E} C_N^k \left(\frac{1}{2}\right)^{N-1} = \left(\frac{1}{2}\right)^{N-1} \sum_{k > E} C_N^k.$$

Dễ dàng nhận thấy rằng:

$$\sum_{k > E} C_N^k \leqslant 2^{N-1},$$

Do đó $T_N < 1$ hay $R_N < \varepsilon$. Mệnh đề được chứng minh.

Mệnh đề 2: Với các điều kiện tương tự như bổ đề 1, ta nhận được

$$R_N < \exp\{-N \times c(\varepsilon)\}, \text{ với } c(\varepsilon) = \ln \frac{1}{2\sqrt{\varepsilon(1-\varepsilon)}}.$$

Chứng minh: Trước hết ta thấy rằng với $0 < \varepsilon < 1/2$

thì $$\ln \frac{1-\varepsilon}{\varepsilon} > 0 \quad \text{và} \quad \ln \frac{1}{2\sqrt{\varepsilon(1-\varepsilon)}} > 0.$$

$\forall \alpha > 0$, $\forall k > E$ ta đều có:

$\exp(\alpha k) \geqslant \exp(\alpha . N/2)$ hay $\exp(-\alpha . N/2 + \alpha . k) \geqslant 1$. Kết hợp với bổ đề 1 ta suy ra:

$$R_N < \exp(-\alpha_n, N/2) \sum_{k > E} C_N^k \varepsilon^k (1-\varepsilon)^{N-k} \exp(\alpha . k) =$$

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

Chọn $\alpha = \ln \frac{1-\varepsilon}{\varepsilon}$ suy ra $R_N \leqslant (2\sqrt{\varepsilon(1-\varepsilon)})^N$. Đặt $c(\varepsilon) = \ln \frac{1}{2\sqrt{\varepsilon(1-\varepsilon)}}$ ta được

$R_N < \exp\{-N.c(\varepsilon)\}$. Mệnh đề được chứng minh. Như vậy, nếu như ta tổ hợp càng nhiều thuật toán thành toán tổ hợp thì hiệu quả của thuật toán tổ hợp càng cao. Phương pháp này được ứng dụng rộng rãi trong trí tuệ nhân tạo và chẩn đoán.

III – ĐÁNH GIÁ XÁC SUẤT SAI LẦM CỦA THUẬT TOÁN NHẬN DẠNG TỔ HỢP TRONG TRƯỜNG HỢP CÁC XÁC SUẤT SAI LẦM CỦA MỖI THUẬT TOÁN KHÁC NHAU

Quy tắc quyết định của thuật toán nhận dạng tổ hợp như sau :

$$D_j(S_l) = \begin{cases} 1, \text{ nếu } \sum_{k=1}^{N} \frac{a_{ij}^k}{\sqrt{\varepsilon_k(1-\varepsilon_k)}} > \frac{1}{2} \sum_{k=1}^{N} \frac{1}{\sqrt{\varepsilon_k(1-\varepsilon_k)}} \\ 0, \text{ ngược lại} \end{cases}$$

Bổ đề 2: Nếu mỗi thuật toán ban đầu có xác suất sai lầm ε_l tương ứng, thì xác suất sai lầm của thuật toán tổ hợp được xác định như sau :

$$R_N = \sum_{\vec{\lambda} \in TA} \prod_{k=1}^{N} \varepsilon_k^{\lambda_k} (1-\varepsilon_k)^{1-\lambda_k}$$

$\vec{\lambda}$ và TA được xác định như trong I.

Chứng minh: Dễ dàng suy ra từ tính độc lập của các thuật toán đã cho. Bổ đề 1 là trường hợp đặc biệt của bổ đề 2 khi $\varepsilon_l = \varepsilon$, $\forall i$.

Mệnh đề 3: Với các điều kiện tương tự như bổ đề 2 thỏa mãn thì

$$R_N \leqslant \frac{4}{N^2} \left(\sum_{k=1}^{N} \sqrt{\varepsilon_k(1-\varepsilon_k)} \right)^2 \Big/ \left(1 - 2 \sum_{k=1}^{N} \tau_k / N \right)^2$$

Chứng minh :

Trước hết ta chứng minh bổ đề sau :

Bổ đề 3: Cho hai dãy số thực $a_1, a_2, \ldots, a_N$ và $b_1, b_2, \ldots, b_N$.

1) Nếu $\forall i, j$; $i > j$, $a_i \leqslant a_j$, $b_i \geqslant b_j$ thì

$$\sum_{i=1}^{N} a_i \sum_{i=1}^{N} b_i \geqslant N \sum_{i=1}^{N} a_i b_i$$

2) Nếu $\forall i, j$ $(a_i - a_j)(b_i - b_j) \leqslant 0$ thì

$$\sum_{i=1}^{N} a_i \sum_{i=1}^{N} b_i \geqslant N \sum_{i=1}^{N} a_i b_i.$$

Chứng minh : 1) chính là bất đẳng thức Trêbusep.

2) Chỉ cần sắp xếp dãy $\{b_i\}$ theo thứ tự tăng dần thì ta cũng nhận được $\{a_i\}$ có thứ tự hoán vị tương ứng với $\{b_i\}$ sẽ giảm dần. Áp dụng 1) ta được 2) – điều phải chứng minh.

Chứng minh mệnh đề:

Từ bổ đề 2 ta suy ra được

$$R_N \leqslant \left(\frac{1}{2} \sum_{k=1}^{N} \frac{1-2\varepsilon_k}{\sqrt{\varepsilon_k(1-\varepsilon_k)}} \right)^{-2} \sum_{\vec{\lambda} \in T\Lambda} \prod_{k=1}^{N} \varepsilon_k^{\lambda_k} (1-\varepsilon_k)^{1-\lambda_k}$$

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

$$\rightarrow R_N \leqslant \left(\frac{1}{2} \sum_{k=1}^{N} \frac{1-2\varepsilon_k}{\sqrt{\varepsilon_k(1-\varepsilon_k)}} \right)^{-2} \sum_{\vec{\lambda} \in \{0,1\}^N} \prod_{k=1}^{N} \varepsilon_x^{\lambda_k} (1-\varepsilon_k)^{1-\lambda_k}$$

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

Dễ dàng thấy rằng:

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

Từ đó ta có bất đẳng thức:

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

Đặt $a_k = \dfrac{1-2\varepsilon_k}{\sqrt{\varepsilon_k(1-\varepsilon_k)}}$, $b_k = \sqrt{\varepsilon_k(1-\varepsilon_k)}$ và áp dụng bổ đề 3 ta có:

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

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

Áp dụng (1) và (2) ta được

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

Nhận xét: – Nếu áp dụng bất đẳng thức Svac cho 2N số thực ta có:

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

Ta thu được $R_N \leqslant \dfrac{4}{N} \dfrac{\bar{\varepsilon}(1-\bar{\varepsilon})}{(1-2\bar{\varepsilon})^2}$; trong đó $\bar{\varepsilon} = \sum_{k=1}^{N} \varepsilon_k$,

Trong trường hợp $\varepsilon_i = \varepsilon$, $\forall i$ ta thu được $R_N \leqslant \dfrac{4}{N} \dfrac{\varepsilon(1-\varepsilon)}{(1-2\varepsilon)^2}$.

Như vậy trong trường hợp các $\varepsilon_i = \varepsilon$, $\forall i$, thì kết quả II mạnh hơn III.

Một số vấn đề còn mở:

– Phát biểu định nghĩa cho thuật toán độc lập.

– Trong trường hợp có định nghĩa về thuật toán độc lập thì hiệu quả của thuật toán tổ hợp từ những thuật toán không độc lập sẽ ra sao? Chúng tôi chân thành cảm ơn tiến sĩ Bạch Hưng Khang và phòng nhận dạng Viện KHTTĐK về sự giúp đỡ nhiệt tình và có những gợi ý quý báu cho bài báo này. Chúng tôi xin cám ơn đồng nghiệp Nguyễn Thanh Thủy trường đại học Bách Khoa Hà Nội, đã có những ý kiến bổ ích cho bài báo.

Nhận ngày 8-3-1986

TÀI LIỆU THAM KHẢO

1. **Hoàng Kiếm**, Some methods improving the efficiency of pattern recognition algorithms. Computer and artificial intelligence, 3(1984), N°4, pp. 347–359, Czeckoslovakia.

2. **Gaillat G., Borne Superieuse** de la probabilité d'erreur avec la gégle de decision majoritaire. Congres AFCET–IRIA, Paris 1978.

3. ЖУРАВЛЁВ Ю И., Об алгебраическом походе к решению задач распознавания или классификации. ВКН: Побл. Кибернетики. Выч. зз. М: Наука, 1978 с. 5–68.

4. РЯЗАНОВ В.В., Комитетный синтез алгоритмов распознавания и класенфикации. Ж. вычисл. Матем. и матем. физ., 1981, Т. 21, И°6. с. 1535–1543.

5. РЯЗАНОВ В.В., О синтезе классифирующих алгоритмов на конечных множествах алгоритмов классификации. Ж. вычисл. Матем. и Матем. физ., Т.22, И°2, с. 429–440.

ABSTRACT

Some results of the efficiencies of the combined pattern recognition algorithm

The paper presents the evalution of the error probability of the combined pattern recognition algorithm based on N. independing algorithms. These main results are follows: if every given algorithm has error probability $0 < \varepsilon < 1/2$ then $R_N < \exp\{-N \cdot c(\varepsilon)\}$, here

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

When all of the given algorithms have error probabili ty $0 < \varepsilon_i < 1/2$ respectively, then the error probability of combined pattern recognition algorithms satisfies

$$R_N \leqslant \frac{1}{N^3}\left(\sum_{i=1}^{N}\sqrt{\varepsilon_i(1-\varepsilon_i)}\right)^2 / \left(1 - 2\sum_{i=1}^{N}\varepsilon_i/N\right)^2,$$