Vol. 22 No. 6 情報処理学会論文誌 Nov. 1981

ショートノート

判別しきい値選定法の一改良†

斎藤泰一†† 山田博三††

文字データを対象にした場合における判別しきい値選定法の改良を行った．文字データの特殊性を実データで示し，判別しきい値選定法の性質を調べ，新しい方法の妥当性について検討している．

1. まえがき

多値図形を二値化する場合その手法は，(1)単純で，(2)対象に依存するパラメータ数が少ないことが望ましい．特に文字データの場合は大量に処理することから，単純かつどのデータに対しても安定したしきい値を与える二値化法を採用したい．

これまでわれわれは非常に単純な全平均濃度レベル$+\alpha$をしきい値として二値化を行ってきた．その対象は，数字，英字，カタカナであったため不自然さはなかった．しかし，漢字，ひらがな混在データに対しては問題がでてきた．それは漢字用にαを決定するとひらがなに対してはαが小さすぎノイズを拾ってしまうのである．理由は，ひらがなは漢字に比べ文字部分の面積が小さく，全平均濃度レベルが低くなるからである．漢字のセット内でも文字部の面積が小さい字が存在し同様のことが言える．

この問題を救い，(1)，(2)の条件を満たす候補として，大津の判別しきい値選定法 (DTSM)[1]が挙げられる．この方法は濃度分布をあるしきい値で二つのクラスに分けたときにクラス間分散とクラス内分散の比が最大になるようなしきい値を見つけるという単純なもので，しかもパラメータがないことが魅力的である．しかし，この方法は濃度分布の歪みが直接しきい値に反映されその悪影響が現われてしまうことがある．

以下，文字データを対象にした場合において，DTSM の短所を補う方法を開発したので，文字データの特殊性を実データで示し，DTSM の性質を調べ，新しい方法の妥当性について検討する．

† An Improvement of the Discriminant Threshold Selection Method by TAIICHI SAITO and HIROMITSU YAMADA (Pattern Processing Section, Information Sciences Division, Electrotechnical Laboratory).

†† 電子技術総合研究所パターン情報部図形処理研究室

2. 記法の定義と判別しきい値選定法 (DTSM)[1]

与えられた図形はLレベルの濃度スケール $S=\{0, 1, \cdots\cdots, L-1\}$ で表現されているものとする．レベル i の画素数を n_i，全画素数を $N=n_0+n_1+\cdots+n_{L-1}$ とする．正規化ヒストグラムとして次式を定義し，これで濃度分布を表現する．

$$p_i=n_i/N, \quad (i\in S,\ p_i\geqq 0,\ \sum_{i=0}^{L-1} p_i=1)$$

図形の全平均濃度レベル，全分散はそれぞれ次式で与えられる．

$$\mu_T=\sum_{i=0}^{L-1} i p_i, \qquad \sigma_T{}^2=\sum_{i=0}^{L-1}(i-\mu_T)^2 p_i$$

今，レベル k をしきい値として，$S_0=\{0, 1, \cdots\cdots, k\}$，$S_1=\{k+1, \cdots\cdots, L-1\}$ に属する画素をそれぞれ2クラス C_0，C_1 に分類しようとする．このとき次の 0，1 次モーメント

$$w(k)=\sum_{i=0}^{k} p_i, \qquad \mu(k)=\sum_{i=0}^{k} i p_i$$

を用いて，各クラスの生起確率は

$$w_0=\sum_{i\in S_0} p_i=w(k), \quad w_1=\sum_{i\in S_1} p_i=1-w(k).$$

また，各クラスの平均レベルは次式で表わされる．

$$\mu_0=\sum_{i\in S_0}\frac{i p_i}{w_0}=\frac{\mu(k)}{w(k)}, \qquad \mu_1=\sum_{i\in S_1}\frac{i p_i}{w_1}=\frac{\mu_T-\mu(k)}{1-w(k)}$$

このとき，$w_0\mu_0+w_1\mu_1=\mu_T$，$w_0+w_1=1$ である．

また，各クラスの分散は

$$\sigma_0{}^2=\sum_{i\in S0}(i-\mu_0)^2 pi/w_0,\ \sigma_1{}^2=\sum_{i\in S_1}(i-\mu_1)^2 pi/w_1.$$

以上の準備から，DTSMは次式のクラス間分散 $\sigma_B{}^2$ とクラス内分散 $\sigma_W{}^2$

$$\sigma_B{}^2=w_0w_1(\mu_0-\mu_1)^2=\frac{[\mu_T w(k)-\mu(k)]^2}{w(k)[1-w(k)]},$$

$$\sigma_W{}^2=w_0\sigma_0{}^2+w_1\sigma_1{}^2$$

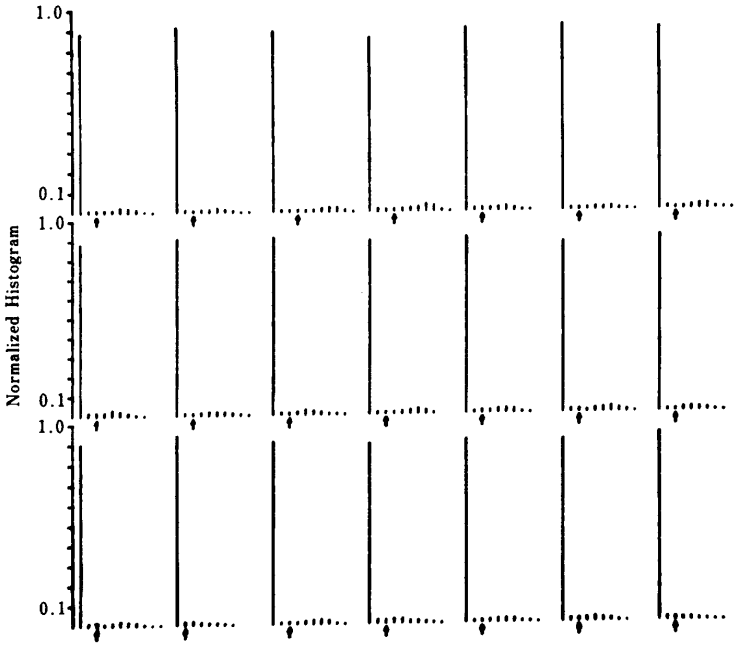

図 1 実データ“愛”の正規化ヒストグラム（21 データ，矢印は k^*）
Fig. 1 Normalized histograms of samples “愛” s (21 samples. Arrows point k^* values).

の比を最大にする k を見つけることになる．$\sigma_W^2+\sigma_B^2=\sigma_T^2$ で σ_T^2 が一定であることから，結局最適しきい値 k^* は σ_B^2 を最大にする k で与えられる．また，レベルをアナログにモデル化した場合にこの最適しきい値は 2 クラスの平均レベル μ_0, μ_1 の中点となる性質[1]から，ディジタル・レベルについてもこの性質を生かし，これをアナログ最適しきい値 $k_a=(\mu_0+\mu_1)/2$ と呼ぶ．これでより精度の高いしきい値が得られる．ここで，$k^*=[k_a]$ である．[ ] は整数部分をとることを意味する．二値化の定義を，レベル i の画素を $i\leqq$(しきい値) のとき C_0 に，$i>$(しきい値) のとき C_1 にクラス分けすることとすれば，k^*, k_a どちらをしきい値にしても結果は同じになる．

3. 新しいしきい値決定法

3.1 文字データの性質

図 1 は手書教育漢字データベース ETL—8[2]の“愛”21 データについてその正規化ヒストグラムを表示したものである．矢印は DTSM の最適しきい値 k^* を示している．この k^* で二値化した図形を**図 3** に示す．

これら文字データの性質を述べると，(1)白地を示す濃度レベルをとる画素数が非常に大きい(2)白地から文字部へ遷移する濃度レベルの分布がなめらかでなく急激に小さくなっている(3)文字部を示す濃度レベルをとる画素数が白地に比べ少なくその分布が平坦に近いことが挙げられる．

3.2 判別しきい値選定法 (DTSM) の性質

3.1 を参考にして濃度分布のモデルを与え，それを変化させたとき DTSM の最適しきい値がどう変わるかを調べる．

濃度レベル数を $L=16$ とし，正規化ヒストグラム $pi(i=0\sim L-1)$ を $p_0=p_1\equiv p_W$, $p_2=p_3=\cdots=p_{L-1}\equiv P_B$ とする(**図 2** (a))．ただし $\sum_{i=0}^{L-1} p_i=1$ より $2p_W+(L-2)p_B=1$ である．この $(L-2)p_B=14\,p_B$ を 1 から小さくしてゆく．つまり白地に対し文字部にあたる濃度レベルの分布の割合を各レベル一様に小さくして行くことを意味する．このとき DTSM のアナログ最適しきい値 k_a の値をプロットしたのが図 2 (b)である．この図より，$14\,p_B$を小さくして行くと k_a は濃度レベル 2〜15 の約 1/3 のレベル（レベル 5.5）に近づき，さらに小さくすると飛ん※でレベル 0.5 に収束する．この飛びを起こす所は，$14\,p_B=0.003$ とかなり小さく現実のデータではほとんどあり得ない状態である．先の実データ（“愛” 21 データ）について文字部の濃度レベルの分布を平担と仮定した場合における文

※ この飛びを起こす原因は，この付近では評価関数 (σ_B^2) が双峰性だからである．また，$14\,p_B>0.4$ で曲線が整数レベル上にのらない理由はモデルの濃度レベルがディジタルなためである．

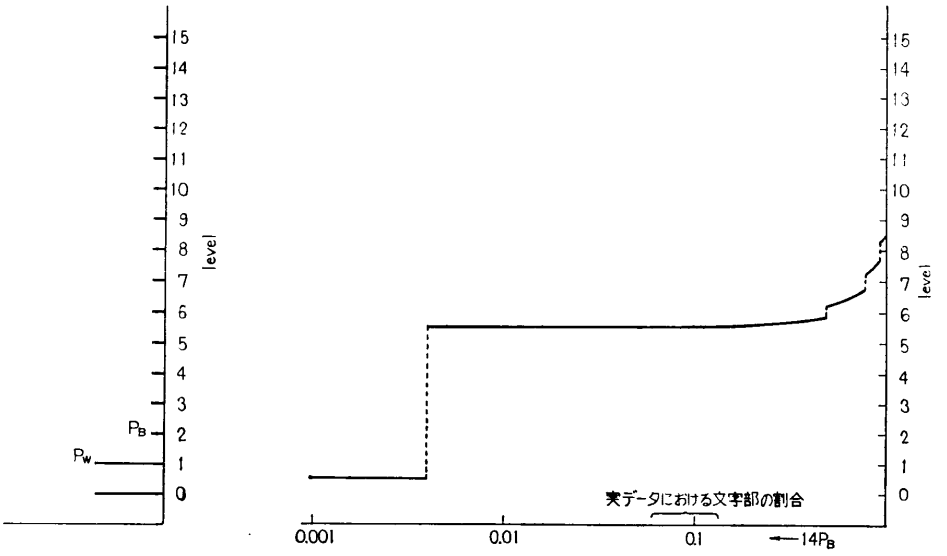

図 2(a) 濃度分布のモデル
Fig. 2(a) A model of a density distribution.

図 2(b) 最適しきい値の変化
Fig. 2(b) Transition of the optimal threshold.

字部の割合の範囲を図中に記す.

このモデルは，特異現象があることを示すためにレベル0と1の2レベルについて p_W を与えたが，$p_0 \equiv p_W$，$p_1 \sim p_{L-1} \equiv p_B$ のようにレベル0だけに p_W を与えた場合は急激な変化は起こらず約 1/3 のレベルに収束するという現象のみが現われる.

このように DTSM では文字部の割合が白地に比べ非常に小さく，しかも文字部にあたる濃度レベルの分布が平坦な場合は，普通平坦な部分内の約 1/3 の濃度レベル以下のしきい値をとり得ないという性質がある.

3.3 新しいしきい値決定法

文字データの性質から，ノイズを拾わない限りできるだけ白地に近いレベルをしきい値としたい．“切れ”をがまんするかノイズを拾うかの二者択一をせまられる最悪の状況においても情報量保存の立場からしきい値は低めにとりたい．DTSM で濃度分布が歪んでいる場合 1/3 のレベル以下のしきい値をとり得ないという性質は，望む所よりしきい値が高めにとられることを意味する．これは図3で“切れ”が目だつことでも分かると思われる.

以上のことから，DTSM の最適しきい値よりいくらか低い所に移動させるために全平均濃度レベル μ_T を導入し，λ をパラメータとして新しいしきい値 T^* を次のように設定した.

$$T^* = \mu_T(1-\lambda) + k_a\lambda, \quad (0 \leqq \lambda \leqq 1).$$

この式は全平均濃度レベル μ_T と DTSM のアナログ最適しきい値 k_a の間の値をとることを意味している.

このしきい値を使って先のデータ“愛”を二値化したものが**図4** ($\lambda=0.25$) である．DTSM の最適しきい値で二値化したもの（図3）に比べ“切れ”が少なくなっていることがわかる.

3.4 新しいしきい値 T^* の性質

(a) T^*の計算は，DTSM を実行することにより k_a と μ_T が求まることから DTSM と同程度の単純さになる.

(b) T^*を変形すると $T^* = \mu_T + (k_a - \mu_T)\lambda$ つまり，$\mu_T + \alpha$ の形をしてその α がデータに依存して変化するものと見ることができる.

(c) ここでは地の部分を白としているが，地の部分が黒い全く逆転した濃度分布をもつデータに対しても T^*がそのまま使える.

(d) 濃度分布が左右対称に近づくにつれ，k_a と μ_T の値が近くなり，濃度分布の歪みが少ないときに悪影響を与えることがない.

(e) もちろん $\lambda=1$ とすれば DTSM そのものになる.

4. あとがき

文字データを対象として，判別しきい値選定法 (DTSM) を全平均濃度レベルで修飾することにより良好な二値化結果を得た．しかし，全平均濃度レベル

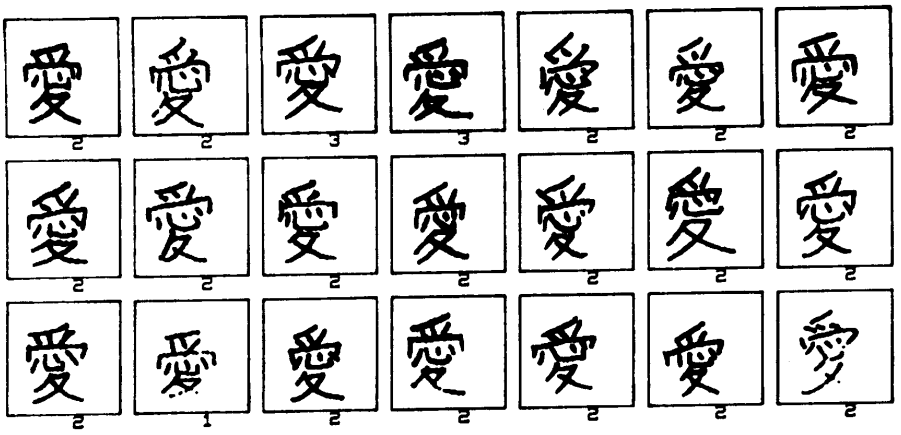

図 3 判別しきい値選定法による二値化データ（数字は k^*）

Fig. 3 Samples dinalized by the Discriminant Threshold Selection Method (The numbers are k^*s).

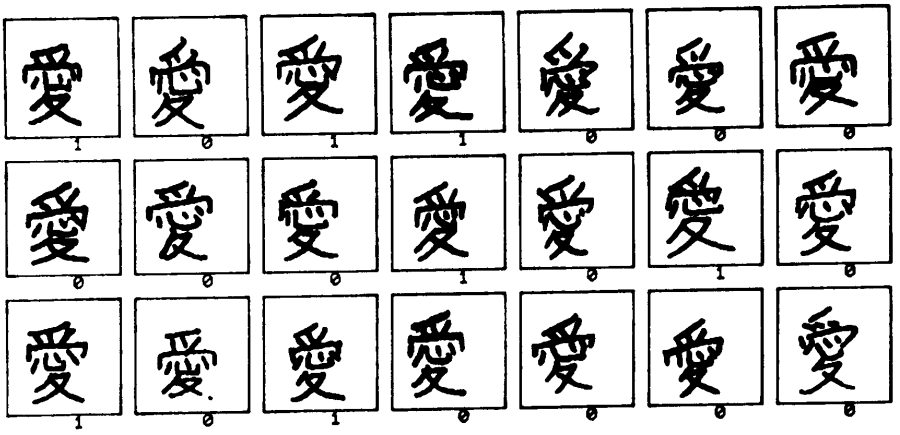

図 4 新しいしきい値による二値化データ（$\lambda=0.25$）

Fig. 4 Samples binalized by the improved method ($\lambda=0.25$).

と DTSM の最適しきい値とを分配する λ なるパラメータを1つ導入せざるを得なかった．今後はこのパラメータをなくすことが課題として挙げられる．そもそも，新しいしきい値を導入した原因が濃度分布の歪みであることから，濃度分布の歪み率など（全平均濃度レベルも含む）により濃度分布を変形し，その後に DTSM を適用する方法が良いかもしれない．このとき，実際に濃度分布を変形せず，最適しきい値を求める途中段階で歪み率なる係数を加えることにより同じ効果が出せるかもしれない．

謝辞 この研究にご支援いただいた当研究所パターン情報部長淵一博氏に感謝いたします．また有益な助言をいただいた当研究室長森俊二氏，数理基礎研究室大津展之氏をはじめ当研究室の方々に感謝いたします．

参考文献

1) 大津展之：判別および最小2乗規準に基づく自動しきい値選定法，信学論，Vol. 63-D, No. 4, pp. 349-356 (1980-4).
2) 森 俊二，山本和彦，山田博三，斎藤泰一：手書教育漢字のデータベースについて，電総研彙，Vol. 43, No. 11, 12, pp. 752-773 (1979-11, 12).

(昭和55年12月12日受付)
(昭和56年4月27日採録)